Lane Image Semantic Segmentation Technology Based on BiSeNetV2 Network

Xiao Hu[1,*], Mingju Chen[1,2]
[1]*School of Automation and Information Engineering, Sichuan University of Science & Engineering, Sichuan, Yibin, Sichuan, China*
[2]*Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Sichuan, Yibin, Sichuan,China*
**Corresponding Author.*

Abstract: With the rapid development of automatic driving technology, lane image semantic segmentation plays an increasingly important role in intelligent transportation systems. In this paper, a lane image semantic segmentation technology based on the BiSeNetV2 network is proposed. First, we describe the dual-branch structure and feature fusion module in the BiSeNetV2 network, and then elaborate on our improvements in the lane image semantic segmentation task. We incorporated the attention mechanism to help the model grasp the overall structure of the image more effectively and enhance the segmentation accuracy. Simultaneously, we introduce depth separable convolution to decrease computational redundancy and simplify the model's complexity. Ultimately, we performed experiments on the Cityscapes dataset, and the results revealed that the proposed algorithm comprises 1.21×10^7 parameters, with an average intersection ratio of 71.4%. At the same time, the network model and algorithm proposed are contrasted with other equally sophisticated techniques. The comparison findings demonstrate that our approach successfully enhances the accuracy and real-time performance of lane image segmentation in comparison to alternative methods.

Keywords: Image Semantic Segmentation; BiSeNetV2 network; Dual-branch Structure; Feature Fusion; Attention Mechanism

1. Introduction

Lane image semantic segmentation has garnered significant interest as a key component of automatic driving systems due to the ongoing advancements in automatic driving technology. The primary objective of this technique is to classify various elements within the lane image, including lane lines, vehicles, pedestrians, and more, into distinct semantic categories. Traditional research on semantic segmentation of lane images mainly depends on traditional image processing techniques, such as edge detection, region growth, and so on. Although these methods are effective in some simple scenarios, they are difficult to meet the challenges in complex scenarios.

In recent years, the advancement of deep learning technology has ushered semantic segmentation into a new phase of growth. In 2015, Long[1] et al. The introduction of Fully Convolutional Networks (FCN) creatively transformed the fully connected layer from the classification network VGG16 into a convolutional layer. Additionally, by implementing skip connections between shallow and deep features, FCN greatly enhanced the results of semantic segmentation. Since the introduction of the FCN method, deep learning has been extensively applied in the area of semantic segmentation, resulting in the emergence of numerous semantic segmentation algorithms that are built on Convolutional Neural Networks (CNN)[2]. Badrinarayanan[3] et al. suggested a complex encoder-decoder architecture using deep convolutional layers through SegNet, in which a pooled index is used to store the position information of pixels in the encoder, and in the decoder, the pooled index of the corresponding encoder is used to perform upsampling, thereby improving the edge segmentation of objects for image segmentation. This technique is capable of categorizing the image on a pixel-by-pixel basis, to realize the

semantic segmentation of the whole image. Subsequently, the DeepLab and DeepLabv2 networks proposed by Chen et al[4]. Employ fully connected conditional random fields (CRFs) for refining the segmentation outcomes, which improves the ability of the model to capture fine edge details. Lin[5] et al. In 2017, RefineNet introduced a multi-path refinement network for performing semantic segmentation on high-resolution images. This method can segment images through multiple paths to achieve higher segmentation accuracy. However, due to the complex network structure, the real-time performance is not good[6].

When extracting image features, the pooling layer and step convolution effectively increase the receptive field is ideal for capturing semantic information, but reduces the image resolution and results in a loss of information about the details of the object. To address the issue of lost detailed information about objects, the above algorithms use codec network structure, skip connection or CRF post-processing methods, but these algorithm structures are still redundant and complex, leading to a substantial rise in the quantity of network parameters[7]. In the feature extraction stage, it can not only effectively obtain rich semantic information, but also reduce the loss of object details, while reducing the number of network parameters as much as possible, To ensure the safety of unmanned driving, which requires the speed of road segmentation is particularly important, so lightweight networks are gradually coming out. BiSeNet proposed by Yu[8] is one of them. BiSeNet uses a dual-branch structure and introduces feature fusion to minimize the computational complexity while ensuring quick processing speed, which is often used in unmanned driving, augmented reality and other fields with high real-time requirements.

Based on the BiSeNet network, this paper proposes a two-branch network model, which obtains the detail information and semantic information of objects through two branches respectively. The shallow network branches are used to retain the detailed information in the image to produce high-resolution characteristics, and the deep network branches are used to downsample to obtain semantic information. The shallow network branch can effectively minimize the loss of detailed information and enhance the accuracy of pixel placement, the deep network branch uses lightweight backbone network downsampling, which can not only extract semantic information, but also decrease the quantity of model parameters and computation. Finally, the feature information obtained by the two branches is effectively fused to further improve the accuracy while ensuring the segmentation speed, and a high-level semantic segmentation model suitable for road scenes is obtained.

2. BiSeNetV2 Network Overview

BiSeNetV2[9] is a neural network model specially designed for real-time scene image segmentation. The two primary branches of the system are the detail branch and the semantic branch. They are tasked with extracting both low-level detailed features and high-level semantic features from the image, respectively, and combine the two through a feature fusion module called bilateral guided aggregation layer, while using an auxiliary segmentation head composed of 1×1 convolution and 3×3 convolution to improve the feature extraction during the training phase. In this way, BiSeNetV2 can maintain high accuracy while maintaining fast operation speed, the network structure diagram depicted in Figure 1 is the one being referred to.

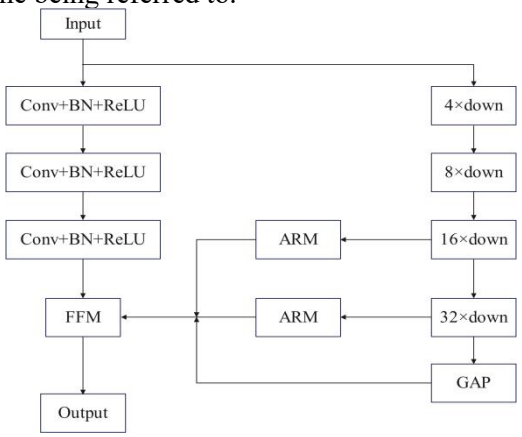

Figure 1. Network Structure of BiseNetV2

2.1 Detail Branch

The main function of the detail branch is to capture the low-level spatial detail information in the image. To achieve this goal, the detailed branch usually has a wide channel and a shallow number of layers. This has the advantage of retaining more spatial information while remaining computationally

efficient. The feature representation of the detail branch has a larger spatial size and a wider channel, which can better encode the spatial detail information.

The detailed branch's network architecture comprises three stages, each incorporating a 3x3 convolutional layer, batch normalization, and an activation function[10]. The initial layer in each stage employs a stride of 2, with the subsequent layers in the same stage maintaining consistent filter quantities and output feature map dimensions. This approach aims to guarantee a reduction in feature map size by half after each stage, thus yielding a final feature map size one-eighth of the original input.

2.2 Semantic Branch

Semantic branches actively engage in extracting high-level semantic information. in the BiseNetV2 network. It has narrow channels and deep layers, which means that it reduces the channel capacity while achieving a lightweight design through a fast downsampling strategy. This design allows the semantic branch to effectively extract high-level semantic features in images while maintaining the lightweight of the network.

The branch uses a lightweight network ResNet18[11] as a backbone network, which can quickly downsample 1/32 of the original graph to obtain rich semantic information, add a global average pool after sampling to further extract and integrate global semantic information, and decrease the size of the feature graph, through global average pooling, so that the model is lighter. At the same time, it can also reduce the use of computing resources[12]. In addition, global average pooling also helps to enhancing the model's generalization ability can decrease its reliance on specific input sizes.

Attention module (ARM): Figure 2 displays the attention module. It compresses the features of each channel into a single numerical value through global average pooling (GAP) and then uses convolution transformation and batch normalization (BN) to extract the feature importance of each channel. These weights are then normalized using the sigmoid function, resulting in an attention weight map between 0 and 1.

This weight map can be used to reweight the original feature map so that useful information gets more attention and unimportant information tends to be ignored. When using the spatial attention module to recalibrate the feature map, for a given input feature map, the calculation and recalibration process of its spatial attention weight can be expressed as:

$$\alpha = \sigma(f^{1\times1}(X)) \quad (1)$$

$$X_{SA} = f_{SA}(X, \alpha) \quad (2)$$

Where: σ is the sigmoid function; $f^{1\times1}$ represents the convolution operation with the convolution kernel size of 1×1; α is the spatial attention weight; (X, α) represents the multiplication of the input feature map and the corresponding spatial attention weight; X_{SA} is the calibrated feature map.

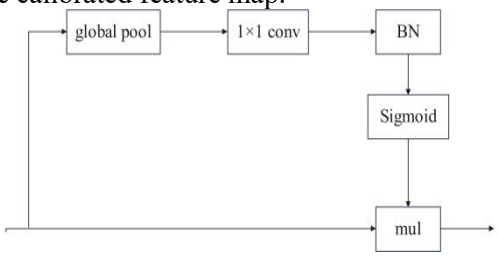

Figure 2. Attention Module

Meanwhile, the semantic branch's feature extraction module utilizes depth-separable convolution to capture the image's high-level features. Depth-separable convolution is a more efficient convolution operation, which decomposes the ordinary convolution into two independent steps: convolution in depth and convolution point by point. Deep convolution uses a small convolution kernel (e.g. 3x3) to conduct a distinct convolution operation for every input channel, which can effectively reduce the number of parameters and provide better local feature extraction capabilities. Then, the point-by-point convolution uses a 1x1 convolution kernel to perform inter-channel linear combination on the output of the depth convolution, to maintain the same dimension of the feature map and greatly improve the accuracy and efficiency of the model while fusing the features of different channels.

Finally, to enhance the accuracy of the segmentation further, the boosting part of the BiSeNetV2 model is used as an intensive training strategy, and auxiliary segmentation heads are added to different positions of semantic branches to carry out additional supervision on the middle output of the model, so that in the training stage, one way to boost the model's accuracy is by improving the

feature representation, specifically during the reasoning stage. These enhanced feature representations can be discarded directly without increasing the inference speed of the model. This strategy can enhance segmentation performance without adding any inference costs.

2.3 Feature Fusion Module

Feature Fusion Module (FFM) usually contains a series of convolutional layers, batch normalization, and activation functions, which help extract and enhance high-level semantic features in images. In BiSeNetV2, the features obtained by the detail branch are mainly low-level features, while the features obtained by the semantic branch are mostly high-level semantic information features. Due to the disparity in levels between spatial and semantic information, the effect of feature fusion directly will be relatively poor. Therefore, this paper uses the channel attention feature fusion module to minimize the distance between the spatial and semantic layer characteristics to enable the efficient integration of various feature levels as depicted in figure3.

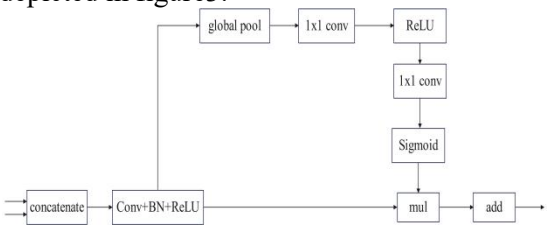

Figure 3. Feature Fusion Module

Firstly, the features at different levels are connected to complete the preliminary fusion to obtain the feature Y, next, the fused features are subject to global average pooling to generate the feature vector Yavgpool that encapsulates the global context, the formula is:

$$Y_{avgpool} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_{(i,j)} \quad (3)$$

In the formula, H and W represent the height and width of the feature map; $Y_{(i,j)}$ represents the element at the (i,j) position;

Then, $Y_{avgpool}$ was subjected to dimensionality reduction by 1×1 convolution channel, the introduction of nonlinear ReLU function, and dimensionality reduction by 1×1 convolution channel, respectively. Then, the convolution result was activated by the sigmoid function to obtain the Favgpool feature, which was finally fused with feature Y to get the ultimate result.

The channel attention feature fusion module introduces the global pooling operation to encompass the worldwide perspective of the characteristics, maximum pooling is employed to gather local feature details from the feature map and enhance the understanding of the feature map's channel significance. [13]. The important information or significant information is adaptively selected from the spatial layer features and the semantic layer features, and the redundant information is suppressed. Finally, the re-weighted features are connected to get the final output, which greatly enhances the network's capability to represent features.

Feature fusion module is very important to enhance the accuracy of semantic segmentation for lane images in real-world scenarios. Assisting the network in gaining a deeper understanding of the semantic information within the image, which is essential for security and reliability in application scenarios such as autonomous driving. Through the effective design and utilization of the feature fusion module, BiSeNetV2 can achieve high-quality semantic segmentation while ensuring real-time performance.

3. Analysis of Experimental Result

3.1 Dataset

Cityscapes is a widely used urban scene segmentation dataset for computer vision tasks, which contains high-resolution images of streets from multiple cities. These images are labeled at the pixel level, identifying different objects and scene categories. The Cityscapes dataset provides 5000 images with fine annotation and 20000 images with rough annotation. In this experiment, the images with fine annotation are used. There are a total of 500 verification images, 2975 training images, and 1525 test images in the dataset. The size of each image is 1024x2048 pixels.

3.2 Evaluation Index

There are many criteria to assess the precision of the algorithm in semantic segmentation. In this paper, the average intersection ratio and the number of parameters of the model are used to evaluate the overall performance of the BiSeNetV2 model.

Intersection over Union (IoU) is a widely used evaluation metric in the field of computer vision. especially in semantic segmentation tasks. Its purpose is to assess the level of agreement between the predicted segmentation and the actual label. The calculation formula of IoU is as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP+FP=FN}} \tag{4}$$

TP (True Positive) indicates the count of accurately predicted pixels, while FP (False Positive) represents the count of pixels inaccurately predicted as targets and FN (False Negative) denotes the count of pixels inaccurately predicted as background. The average intersection over union ratio for all classes is referred to as the mean intersection over union ratio (mIOU). A higher mIOU value indicates a greater overlap between the model's prediction and the actual label, leading to better performance.

Params typically represent the overall count of parameters in a model, serving as a crucial metric for assessing the model's scale. The total number of parameters reflects the number of weights and biases that need to be learned and updated during model training. In general, the fewer parameters and the more lightweight the model, the lower the computational cost and storage requirements, but it may also affect the expressiveness and performance of the model.

3.3 Result Analysis

The experimental environment in this paper is Pytorch1.11.0, Cuda11.1, Cudnn8, minicoda Python3.8.8, and Ubuntu18.04; the segmentation effect of the lane image based on the BiSeNetV2 algorithm on the Cityscapes data set in this experiment is shown in Figure 4.

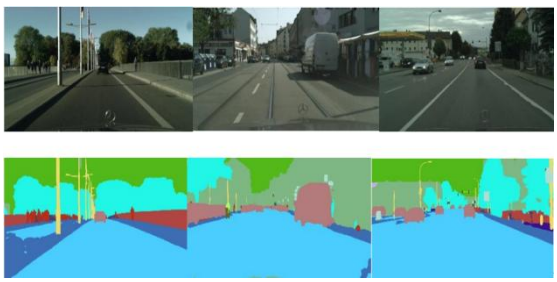

Figure 4. Segmentation Result

The experimental results above demonstrate that the real-time image semantic segmentation algorithm, which is based on the BiSeNeV2 algorithm proposed in this paper, can effectively achieve semantic segmentation of input images in road traffic scenes, and this algorithm has good segmentation performance, and the accuracy of semantic segmentation has been significantly improved.

To better reflect the effectiveness of the double-branch structure semantic segmentation algorithm proposed in this paper, this paper selects SegNet[14], DeepLab, FCN-8s, RefineNet, and other algorithms to compare their performance with the algorithm in this paper. The comparison results are displayed in Table 1 using the Cityscapes dataset.

Table 1 Segmentation Accuracy of Different Algorithms

Algorithm	Backbone	mIOU%	Parameters/10^6
SegNet	VGG16	55.6	29.8
DeepLab	VGG16	62.3	38.3
FCN-8s	VGG16	64.5	145.7
RefineNet	ResNet101	72.0	119.0
BiSeNetV2	ResNet18	71.4	12.1

It can be seen from Table 1 that the proposed BiSeNetV2 algorithm improves the average intersection ratio compared with most other algorithms, indicating that the detail information retained by the detail branch improves the accuracy of pixel positioning and contributes to the segmentation of the edge contour of the target object, thus achieving a better segmentation effect. The RefineNet algorithm achieves a higher average intersection and union ratio, partly because it uses the ResNet101 algorithm with a more complex structure as the backbone network, and its feature extraction ability will be stronger; partly because it uses methods such as skip connection or CRF post-processing in the network, which greatly addresses the issue of lost detailed information resulting from the down-sampling process. However, this algorithm has a significant amount of variables. The parameters of RefineNet are close to ten times the algorithm that is suggested in this paper, and its computational cost and storage requirements will be relatively high. The SegNet algorithm in Table 1 has fewer parameters, but its accuracy is low. Compared with the algorithm proposed in this paper, the quantity of parameters has increased, but the average intersection ratio has increased by 15.8%. To sum up, the algorithm in this paper achieves a better segmentation effect with fewer parameters.

4. Summary

In this paper, a semantic segmentation technology of lane images based on the BiSeNetV2 network is proposed. By introducing attention mechanism and depth-separable convolution, the algorithm ensures high accuracy while reducing the amount of computation and model complexity as much as possible, to enhance the effectiveness and versatility of the model. Through the experimental verification, BiSeNetV2 achieves an excellent performance of 71. 4% average cross-union ratio on the Cityscapes data set. In the future, the research work will further explore how to apply the algorithm to the actual scene, in order to improve its robustness and practicability in complex environments.

Acknowledgments

This research was funded by the Natural Science Foundation of Sichuan (grant number 2023NSFSC1987),

References

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[2] Yu C, Gao C, Wang J, et al. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 2021, 129: 3051-3068.

[3] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 2017, 39 (12): 2481-2495.

[4] Chen L C, Papandreou G, Kokkinos I, 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, 40 (4): 834-848.

[5] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.

[6] WANG Longfei, YAN Chunman. Overview of semantic segmentation of road scenes. Advances in Lasers and Optoelectronics, 2021, 58(12): 44-66.

[7] Mingju Chen, Hongyang Li, Hongming Peng, Xingzhong Xiong, Ning Long. HPCDNet: Hybrid position coding and dual-frquency domain transform network for low-light image enhancement. Mathematical Biosciences and Engineering, 2024, 21 (2): 1917-1937.

[8] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.

[9] Yu C, Gao C, Wang J, et al. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 2021, 129: 3051-3068.

[10] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR. org, 2015: 448 − 456.

[11] Chen M, Yi S, Lan Z, et al. An Efficient Image Deblurring Network with a Hybrid Architecture. Sensors, 2023.

[12] Xia T H, Tan M, Li J H, et al. Establish a normal fetal lung gestational age grading model and explore the potential value of deep learning algorithms in fetal lung maturity evaluation. Chinese Medical Journal, 2021, 134.

[13] Wang F, Luo X, Wang Q, et al. Aerial-BiSeNet: A real-time semantic segmentation network for high resolution aerial imagery. Chinese Journal of Aeronautics, 2021.

[14] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (12): 2481 – 2495.

