Computer Science and Information Systems 20(4):1749–1770 https://doi.org/10.2298/CSIS221205029Z

LUN-BiSeNetV2: A Lightweight Unstructured Network Based on BiSeNetV2 for Road Scene Segmentation

Yachao Zhang[1] and Min Zhang[2]

[1] School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology
Zhengzhou 450064, China
zhangychaoc@163.com
[2] Zhengzhou Commercial Technician College
Zhengzhou 450064, China
zhangminmin@126.com

Abstract. With the continuous introduction of automatic driving technology, the research of road scene segmentation algorithm in machine vision has become very important. In traditional methods, most researchers use machine learning methods to segment thresholds. However, the introduction of deep learning in recent years makes convolutional neural networks widely used in this field. Aiming at the problem that the traditional threshold segmentation method is difficult to effectively extract the threshold value of road image in multiple scenes and the serious problem of over-segmentation caused by deep neural network training data directly, this paper proposes a road scene segmentation method based on a lightweight unstructured network based on BiSeNetV2. The network contains backbone segmentation network and BiSeNetV2 network. The Mobilenetv2 network is used in the backbone network to replace the Xception feature extraction network in the decoder. In addition, grouping convolution is used to replace common convolution in Mobilenetv2 network. And it selects the batch specification layer to reduce the number of parameters, without affecting the accuracy and improving the efficiency of segmentation. At the same time, due to the relatively fixed distribution position of unstructured roads in the image, attention mechanism is introduced to process advanced semantic features, so as to improve the sensitivity and accuracy of the network. The BiSeNetV2 network enhances the dominant relationship between channel features by adding a compression excitation module based on channel attention mechanism after the detail branch, so as to perceive key areas and highlight local features. The lightweight feature pyramid attention mechanism is used to optimize semantic branches, improve the feature integration between contexts, extract high-level road semantic information more efficiently and retain spatial location information to the maximum extent. Finally, local semantic features and high-level semantic features are fused to improve the effect of unstructured road detection. The experiment is trained on the open data set. The results show that compared with other state-of-the-art networks, the accuracy and real-time performance of proposed LUN-BiSeNetV2 in this paper are good, and the false segmentation and edge clarity are better. Compared with the classical algorithm, the average intersection is improved by 2.2% compared with mIoU, the average pixel accuracy is improved by 7.6%, and the frame rate is improved by 24.5%.

Keywords: Road Scene Segmentation, BiSeNetV2, lightweight unstructured network, attention mechanism.

1. Introduction

Image segmentation in driverless cars has become a hot topic in the society. Driverless driving is defined as the realization of highly autonomous driving behavior through environmental perception: starting, braking, lane line tracking, lane change, collision avoidance, parking, etc. Road scene image segmentation plays an important role in this technology [1]. It is of great significance for the development of unmanned driving technology to study how to obtain efficient scene segmentation image in complex scene and serious noise environment [2].

The traditional road segmentation method is based on binocular stereo vision and motion index. Rajendar et al. [3] proposed pedestrian detection based on binocular stereo vision and support vector machine (SVM) algorithm, and determined the coordinate position of moving target by threshold segmentation. In view of the diversity of motion indicators, Raipuria et al. [4] segmented the road with multiple motion indicators such as projected surface direction, object height and feature tracking density. However, the above methods have high requirements on computing resources. In view of the practical requirements of unmanned driving at the present stage, a more simple method with lower resource occupancy rate is needed.

Recently, deep learning has been gradually introduced into road scene segmentation. Teng et al. [5] studied on intelligent vehicle steering based on end-to-end deep learning. Good road feature coding was obtained by pre-training self-encoding. In recent years, AI technology has attracted the attention from many scholars [6]. With the continuous research and development of GPU parallel running, computing acceleration, storage space compression and other technologies, the environmental requirements of excessive data and computation have gradually ceased to be the limitation. Convolutional neural network (CNN) has become a research hotspot of the public again and has been widely used. Kong et al. [7] learned high-order features in the scene based on deep learning algorithm to achieve road scene segmentation. Although the calculation intensity was reduced to a certain extent, the problem of over-segmentation in some complex scenes still existed. Abdalla et al. [8] proposed a method to measure feature similarity in source-target scenes with feature auto-encoder based on the feature automatic extraction capability of deep convolutional neural network (DCNN) depth structure for complex scenes. Zhu et al. [9] proposed image object segmentation by integrating T-node cues, considering that image boundaries could be effectively maintained in the process of image segmentation. However, the segmentation accuracy of these algorithms for road signs, vehicles and pedestrians has not reached the ideal result. The segmentation phenomenon often occurs on the road surface in cloudy and rainy days, snowy days and high temperature days.

Wang et al. [10] used the parallel linear edges of lanes in structured traffic environment as the basis for the detection of passable areas, and proposed a recognition method based on gray features. In order to reduce the impact of light changes and shadows on detection results, Yuji et al. [11] proposed a segmentation method based on lane marker lines by taking advantage of the fact that most edge points towards the expansion center of road boundaries were located at road boundaries. However, the above methods are not applicable to unstructured road detection. The reason is that unstructured roads lack identifiable lane lines, clear road edges and large background differences [12], making it difficult to complete the detection task. Fully Convolutional Networks (FCN) [13] realized end-to-end semantic segmentation, which greatly improved the generalization

ability of convolutional neural networks. In order to improve the segmentation accuracy, scholars have made a series of improvements based on FCN. SegNet [14] improved the up-sampling process by using the recorded maximum pooled index for up-sampling processing, which greatly improved the segmentation accuracy. DeepLabv1 [15] optimized the boundary details by processing the segmentation results using Conditional Random Field (CRF)[16]. DeepLabv2 used dilated convolution to replace the up-sampling method in the original Deeplabv1, and proposed the Atrous Spatial Pyramid Pooling (ASPP) module, which could reduce the number of parameters and increase the accuracy. DeepLabv3 [17] optimized the hollow space pyramid module to better capture multi-scale information. DeepLab3 introduced an encoder-decoder structure based on DeepLabv3 to better integrate low-level semantic features with high-level semantic features. PSPNet [18] proposed pyramid pool module, which could make full use of context information. UNet [19] proposed the U-shaped structure, which could use a smaller number of data sets without reducing the segmentation accuracy.

Since unstructured roads are characterized by many features, wide distribution range, large regional differences, and sensitivity to environmental factors [20]. If the above semantic segmentation network is used for identification, not only a large number of unstructured road data sets with rich types are needed, but also targeted modifications to the network structure are required. Otherwise, too deep network will lead to poor real-time performance and over-fitting phenomenon. DeepLabv3+ network with high segmentation accuracy, strong robustness and relatively simple structure is selected for improvement in this paper. The specific contributions of this paper are as follows.

1. The backbone Xception [21] in DeepLabv3+ network is replaced by MobileNetv2 [22], which can reduce the network structure, improve the feature extraction speed, and prevent over-fitting phenomenon.
2. In MobileNetv2 and dilated space pyramid module, grouping convolution [23] is used to replace traditional convolution, and batch specification layer is deleted, so as to reduce the number of parameters and computation without affecting the accuracy, and improve the segmentation efficiency.
3. The attention module is added to the dilated space pyramid module to improve the speed and accuracy of network recognition.
4. The BiSeNetV2 network enhances the dominant relationship between channel features by adding a compression excitation module based on channel attention mechanism after the detail branch, so as to perceive key areas and highlight local features.

The remaining paper is managed as follows: Section 2 presents the related works in detail. Section 3 discusses the proposed road scene segmentation. Then, in Section 4, experiments and results are provided. Lastly, the paper is concluded in Section 5.

2. Related Works

With the popularization of intelligent transportation and the rapid development of unmanned driving technology, high-precision detection and segmentation of multi-class targets in road scenes based on computer vision is becoming increasingly important. Currently, road scene segmentation based on deep learning method is the most widely used. In mainstream segmentation networks, many factors affect the robustness and richness

of features, and there are various methods to optimize feature performance, including multi-scale feature extraction, spatial context, attention mechanism, etc. One of the novel methods is to train features against interference.

The mainstream road segmentation networks start from the fully convolutional neural networks (FCN). Deconvolution was first proposed in FCN to replace the fully connected structure in convolutional neural network. The fully connected layer was removed, the ability of sharing convolution features was improved, and the prediction of road position pixel level was realized. However, with the increase of network pooling times, the problem of feature loss in encoder structure becomes more and more serious. In order to deal with the problem, SegNet was proposed to improve the image quality of up-sampling by introducing anti-pooling layer. Fan et al. [24] used the residual module on the basis of FCN, which could deepen the depth of the network and increase the propagation path of feature information at the same time, thus improving the detection accuracy and convergence speed of the network respectively. Wang et al. [25] put forward a pyramid pooling module in PSPNet, which used multi-scale pooling operation to output feature graphs of four sizes and carry out convolution operation to obtain image features under different receptive fields. In reference [26], a pyramid pool module of void space was introduced into DeepLabV3, and void convolution with different expansion multiples was used to obtain information of different sensitivity fields, which increased the diversity of extracted features. Kazerouni et al. [27] proposed a U-shaped network structure, which constructed multiple skip connection (SC) between encoder and decoder to realize the combination of low-level features and high-level features, reduce the loss of features during down-sampling, and improve the accuracy of segmentation results.

The above semantic segmentation network improves the segmentation accuracy of the target to a certain extent, and obtains good extraction effect. But it is slightly insufficient in the ability of anti-interference, which may come from the input image or from the features extracted by the network. Xu et al. [28] proposed the denoising convolutional neural network. Firstly, Gaussian noise was artificially added to the input image, and the convolutional network with residual structure was used to remove the noise, so as to achieve the denoising function. However, these noises were artificially added and natural real noise image had a certain difference. In order to simulate the possible interference information more truly, Ding et al. [29] proposed a pseudo-feature generator and linearly superimposed the generated pseudo-features into the intermediate feature layer of the detection network. Through the training method of generating antagonism [30], the constantly changing pseudo-features were obtained, so that the pseudo-features were similar to the real features. However, the detection network needed to overcome these feature interference and obtain the capability of anti-feature interference, which improved the accuracy of road detection. This training method of introducing confrontation to the encoder of detection network belongs to self-supervised learning [31], which indirectly improves the ability of encoder to filter interference by combining two supervised learning methods in training.

The advantage of self-supervised learning compared with supervised learning is that it does not have to rely on a large number of human-annotated labels. In recent years, the contrast learning method [32] has further improved self-supervised learning. The measure function is used to enhance the positive correlation and reduce the negative correlation, so as to improve the ability of the network to distinguish between positive and negative

samples. The contrast learning method pays more attention to the ability of learning feature extraction, and can achieve higher detection accuracy by fine-tuning training with a small amount of labeled data in specific tasks, improve the generalization performance of the network, and effectively reduce the phenomenon of over-fitting.

Since there is a certain correlation between global information and local information in the road scene. For example, the lane often appears in the middle of the image, it learns the attribution relationship between the them that is helpful to improve the prediction ability of the network. Zhu et al. [33] established the attribution relationship between global information and local information by maximizing mutual information between global and local features. Therefore, this paper proposes a novel road scene segmentation method based on a lightweight unstructured network and BiSeNetV2 to improve the robustness of the overall network.

3. Proposed Road Scene Segmentation

DeepLabv3+ network has good semantic segmentation, simple structure and fast segmentation speed. Because the original Xception network will appear over-fitting phenomenon when performing the tasks in this paper, and the speed cannot meet the real-time requirement, so the backbone is replaced by the lightweight transformation of MobileNetv2 network, which avoids the phenomenon of over-fitting phenomenon in the network depth and improves the efficiency of feature extraction. In order to improve the segmentation efficiency, the dilated space pyramid pool structure is improved by lightweight. At the same time, in order to improve the decoder's sensitivity to the target region, the attention module is introduced after the hollow space pyramid structure of the encoder. Then, the BiSeNetV2 network is used to improve the decoder, and the high-level and low-level feature graphs are fused and the pixels are classified. The complete network structure is shown in Figure 1.

3.1. Grouping Convolution

In order to reduce the number of parameters and improve the network efficiency, this paper, inspired by reference [34], uses grouping convolution to replace the common convolution operation in MobileNetv2 and spatial pyramid structure. Grouping convolution operation refers to grouping feature graphs and then convolving them when convolving feature graphs. The principle is shown in Figure 2. For an un-grouped network, the input feature graph size is $C \times W \times H$, the number of groups of convolution kernel is N, and the size of convolution kernel is $C \times K \times K$. To output N groups of feature graphs, it needs to learn $C \times K \times K \times N$ parameters. If the feature graph is divided into G groups, only $(C/G) \times K \times K \times N$ parameters need to be learned, and the total number of parameters is reduced to $1/G$. At the same time, the grouping convolution can also be regarded as making an dropout on the original feature graph, which has the regularization effect and avoids the phenomenon of over-fitting.

3.2. Lightweight Backbone

In this paper, the backbone of the new network adopts the structure of MobiletNetv2, and the specific structure is shown in Table 1.

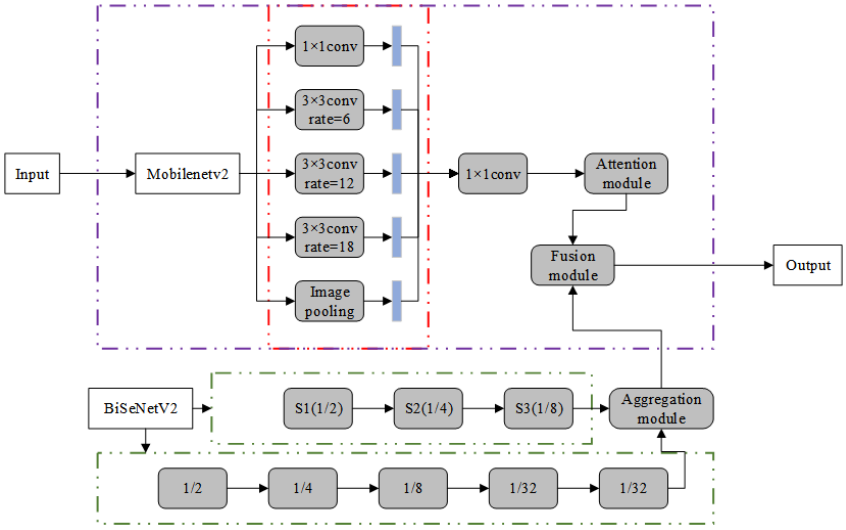

Fig. 1. The overall of proposed model framework

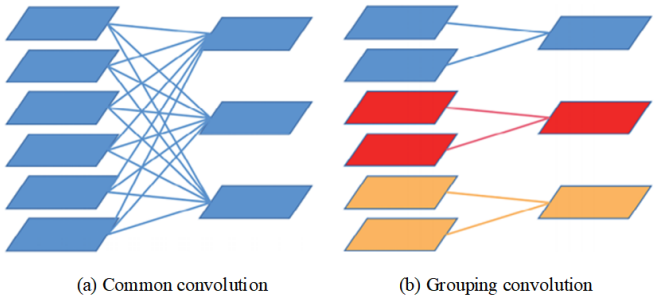

Fig. 2. The schematic diagram of groups convolution

Table 1. Parameters in the main backbone network

Input size ($H^2 \times C$)	Operator	t	c	n	s
$512^2 \times 3$	Conv2d	none	32	1	2
$256^2 \times 32$	Lite-bottleneck	6	16	1	1
$256^2 \times 16$	Lite-bottleneck	6	24	2	2
$128^2 \times 24$	Lite-bottleneck	6	32	3	2
$64^2 \times 32$	Lite-bottleneck	6	64	4	2
$32^2 \times 642$	Lite-bottleneck	6	96	3	1
$32^2 \times 96$	Lite-bottleneck	6	160	3	2
$16^2 \times 160$	Lite-bottleneck	6	320	1	1

In Table 1, H^2 is the number of pixels in the input image. C is the number of input channels. t is the multiplication factor of the input channel. n The number of repetitions of Lite-bottleneck of this size. s is the convolution step length of each Lite-bottleneck module.

In this structure, Lite-bottleneck is a bottleneck that is a lightweight improvement over its previous bottleneck. Lite-bottleneck has the same mechanism as its original bottleneck, and it can be divided into increasing dimension layer, convolutional layer, and dimension reduction layer. First, 1×1 convolution in the increasing dimension layer is used to raise the spatial dimension of the feature graph. Then, the feature of each channel feature graph is extracted by the depth separable convolution of 3×3 in the convolutional layer. Finally, the dimension is reduced by 1×1 convolution. Among them, the activation function of "dimension enhancement layer" and "convolution layer" is ReLU function. In order to prevent ReLU function from destroying the compressed features, the activation function after "dimension reduction layer" uses Linear function, as shown in Figure 3. Lite-bottleneck replaces the common convolution in "dimension raising layer" and "dimension reducing layer" by a block convolution with a bottleneck number of 2, which reduces the number of parameters. At the same time, because grouping convolution prevents over-fitting in shallow networks, it removes batch specification layers in Lite-bottleneck to reduce the computation amount. In the process of feature extraction, the feature map output with the size of 1282×24 is input into the decoder as low-level semantic features, and the feature map is input into the hollow space pyramid pool structure. The specific algorithm of network is shown in **Algorithm 1**. If the input image is bottleneck X and has a total of N layers, the following operation procedure occurs.

Algorithm 1 Lightweight MobileNetv2 algorithm

Input: X

$X_t = Conv(X)$
$X_R = ReLU(X_c)$
for $n = 1$ to N **do**
  $X_1 = ExpandedConv(X_n, groups = 2, noBN)$
  $X_2 = DepthwiseConv(X_1)$
  $X_3 = ProjectConv(X_2, groups = 2, noBN)$
end for
if $n == 2$ **then**
  $X_{LowLevel-Feature} = X_n$
end if
Output: $X_{LowLevel-Feature}, X_n$

Compared with the original Xception of DeepLabv3+, the new structure greatly reduces the network depth and the number of parameters, which is more suitable for the real-time unstructured road detection task in this paper.

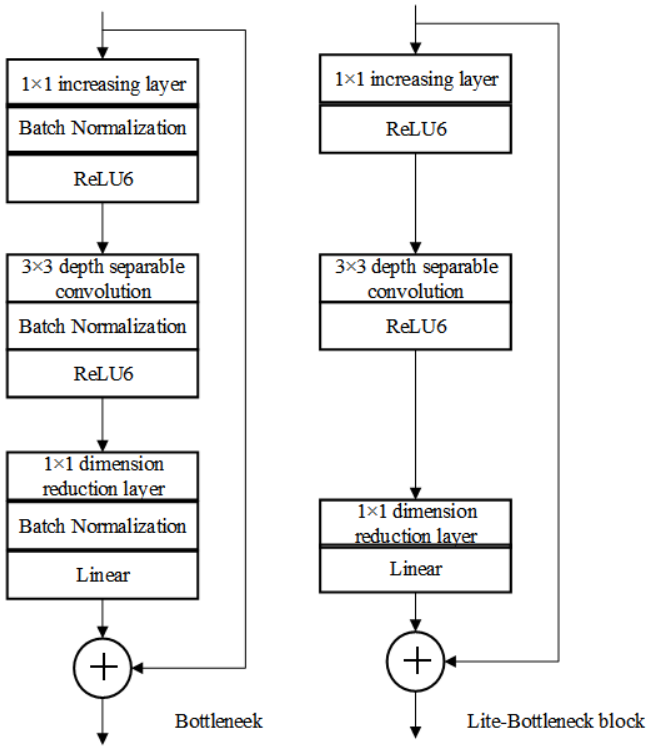

Fig. 3. Structure comparison of original bottleneck and Lite bottleneck

3.3. A lightweight dilated space pyramid pool module integrating attention mechanism

In order to enlarge the receptive field and extract multi-scale feature maps, DeepLabv3+ network adopts dilated space pyramid structure. The convolution with different dilated rates is used, and the large feature maps of different scales are extracted with the same number of parameters, which can better detect large targets like roads. At the same time, multi-scale context information can be captured to locate the target more accurately. After the image is transformed into a space pyramid pool, the feature map with unified dimensions and sufficient information is output. In order to accelerate the segmentation process and realize the function of real-time detection, a lightweight dilated space pyramid pool module is proposed in this paper. The dilated convolution is replaced by the dilated grouping convolution with group number of 2, which can reduce the number of parameters and act as the batch specification layer. Then it deletes the batch specification layer to reduce computation. The feature graphs of the lightweight void space pyramid pooling module are fused together, and the number of channels is adjusted by 1×1 convolution, and then the attention module is passed.

By assigning different weights to pixels, the attention module can adjust the attention focus of the whole network and improve the recognition effect and efficiency. The attention module in this paper is composed of channel attention module and space attention module in series. The channel attention module can focus on the meaningful part of the input feature, while the spatial attention module can obtain the global information in the scene [35].

The structure of the channel attention module is shown in Figure 4. After global average pooling and global maximum pooling, two groups of 1×1 feature maps X^c_{avg} and X^c_{max} are obtained for input feature X, which are then sent into two layers in fully connected neural networks respectively. The two-layer fully connected neural network shares parameters for the two groups of feature graphs, and then adds the two groups of feature graphs to get the weight coefficient between 0 and 1 by Sigmoid function, and multiplies it with the feature graphs to get the optimized feature graph. W_0 and W_1 are used to represent the two parameters of the shared network hiding layer (SharedMLP). The Sigmoid function is represented by σ. The channel attention can be calculated by the following formula.

$$M_c(X) = \sigma(W_1(W_0(X^c_{avg})) + W_1(W_0(X^c_{max}))) \tag{1}$$

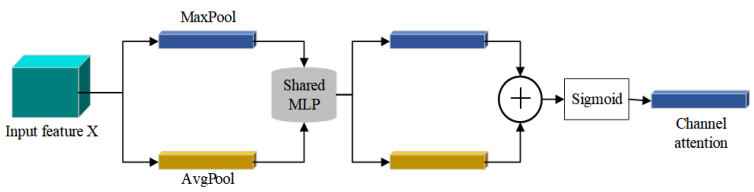

Fig. 4. Channel attention module

The structure of spatial attention module is shown in Figure 5. The optimized feature graph X' by the channel attention module is input. First, after maximum pooling and average pooling with only one channel dimension, the two obtained feature graphs X^s_{avg} and X^s_{max} are spliced together. After another convolution layer, it adopts a 7×7 convolution kernel for convolution operation. At the same time, it keeps the feature map size unchanged. The obtained feature graph is generated by using the Sigmoid function to generate the spatial weight coefficient, which is multiplied with the input feature graph to obtain the final feature graph. The calculation formula is as follows.

$$M_s(X') = \sigma(f^{7\times7}([AvgPool(X'); MaxPool(X')])) = \sigma(f^{7\times7}([X^s_{avg}; X^s_{max}])) \quad (2)$$

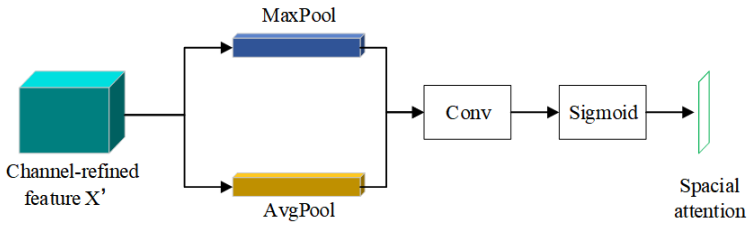

Fig. 5. Spacial attention module

3.4. Detail Branch Network Model Design

In road detection, due to the complexity and importance of road edge details, the network model in this paper uses detail branch with shallow high-channel convolution structure to extract road features. The shallow convolution structure can extract the detailed information, reduce the memory access cost, and improve the model reasoning speed. A higher number of channels can encode a variety of local information. This branch uses S1, S2 and S3 to represent the three sub-sampling stages, as shown in Table 2.

Table 2. Structure of Detail Branch

Stage	Operation	Kernel	Size	repetition	Output size
S1	Conv2D	$64 \times 3 \times 3$	2	1	256×256
S1	Conv2D	$64 \times 3 \times 3$	1	1	256×256
S2	Conv2D	$64 \times 3 \times 3$	2	1	128×128
S2	Conv2D	$64 \times 3 \times 3$	1	2	128×128
S3	Conv2D	$128 \times 3 \times 3$	2	1	64×64
S3	Conv2D	$128 \times 3 \times 3$	1	2	64×64

Each stage consists of 2 to 3 Conv2D convolution blocks, wherein Conv2D operation is shown in Equation (3).

$$O = g_s^{3\times3} = \delta(BN(f_s^{3\times3}(I))) \tag{3}$$

Where, O is the output feature map of Conv2D. $g_s^{3\times3}$ represents Conv2D with convolution kernel of 3×3 and step size of s. I is the input feature map of Conv2D. $f_s^{3\times3}$ is the convolution operation with convolution kernel 3×3 and step size s. BN is batch normalization. δ is the ReLu activation function. After three stages of Conv2D, the final output feature map extracted by the branch is $1/8$ of the original input image.

In road detection, although pooling operation is not carried out in this branch, convolution processing with step size of 2 is used to reduce dimension and reduce the size of the output feature map. In addition, compared with pooling operation, this dimension reduction method can retain more local road feature information through parameter fitting, and the shallow structure does not increase the complexity of the network too much, thus improving the speed of road feature detection.

3.5. Semantic branch network model design

When extracting features from roads, semantic branches are designed to extract advanced semantic features from roads. It is mainly divided into three stages S1, S2 and S3, which represent three stages of Stem block, gather-and-expansion (GE) block and context embedding (CE) block respectively, as shown in Table 3. Extracting high-level semantic information requires a large receptive field, so this branch adopts fast down-sampling strategy to improve the expression level of road features, rapidly expand receptive field, and uses global average pooling to obtain context relations.

Table 3. Structure of Semantic Branch

Stage	Operation	Kernel	Size	repetition	Output size
S1	Stem	$16 \times 3 \times 3$	4	1	128×128
S2	GE	$32 \times 3 \times 3$	2	1	64×64
S2	GE	$32 \times 3 \times 3$	1	1	64×64
S2	GE	$64 \times 3 \times 3$	2	1	64×64
S2	GE	$64 \times 3 \times 3$	1	1	32×32
S2	GE	$128 \times 3 \times 3$	2	1	32×32
S2	GE	$128 \times 3 \times 3$	1	3	16×16
S3	CE	$128 \times 3 \times 3$	1	1	16×16

Stem blocks are adopted in S1 stage, and the structure is shown in Figure 6. The module first convolves the original input image, down-samples the original image, and then uses two different down-sampling methods to reduce the feature representation. Finally, the output features of the two branches are connected as outputs, and a feature graph of 1/4 size is obtained. Due to the using of Stem modules for convolution and pooling operations to narrow the feature map, the structure has efficient computational cost and effective road feature representation capability.

In phase S2, multiple GE blocks are used. This stage belongs to the semantic integration stage of semantic branches, and the structure of each GE block is shown in Figure 7.

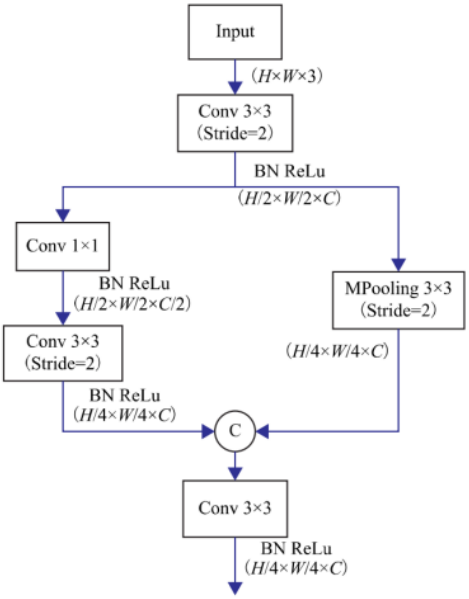

Fig. 6. Architecture of Stem Block

In this module, two 3×3 depth separable convolutions are independently used on each channel for feature extraction. Compared with one 5×5 depth separable convolution, they have the same receptive field. However, the complexity of the module is reduced.

The CE block is used in the S3 phase to integrate contextual information, and the structure is shown in Figure 8. In this module, global average pooling and residual connection are used to effectively extract global context information, so as to achieve the integration of road context information of GE layer. However, the structure of residual-linked will slow down the overall reasoning speed of the model.

3.6. Design of aggregation network model

After the details branch and semantic branch extract the features of different levels of the road, the extracted features need to be fused in the aggregation module to improve the accuracy of feature extraction. The network model design of the aggregation module is shown in Figure 9.

Since the feature graph of semantic branch output is 1/4 of the feature graph size of detail branch output, the detail branch is divided into two parts in this paper. One part uses conventional convolution with step size 2 and convolution kernel size 3×3 and 3×3 average pooling operation with step size 2, and the other part uses depth-separable convolution and 1×1 convolution. The semantic branch is also divided into two parts for processing. One part only uses conventional convolution and four up-sampling operation, while the other uses deep separable convolution with convolution kernel of 3×3 and convolution operation of 1×1. The final detail branch and semantic branch are divided into four branches, and the fusion method is shown in Formula (4).

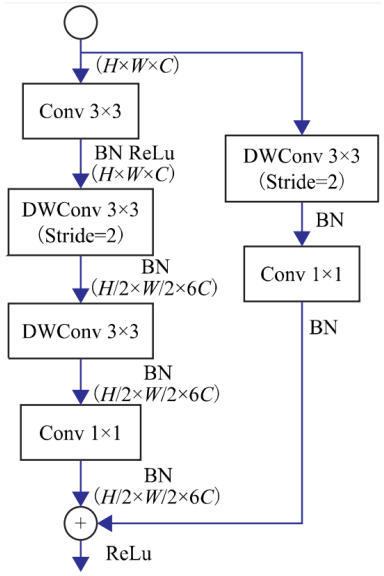

Fig. 7. Architecture of GE Block

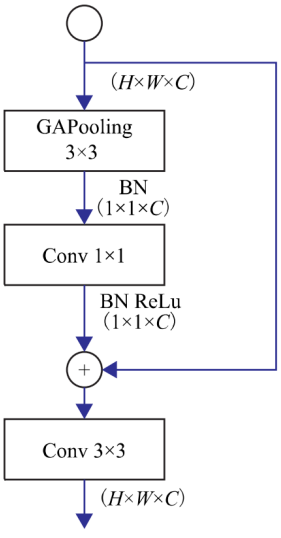

Fig. 8. Architecture of CE Block

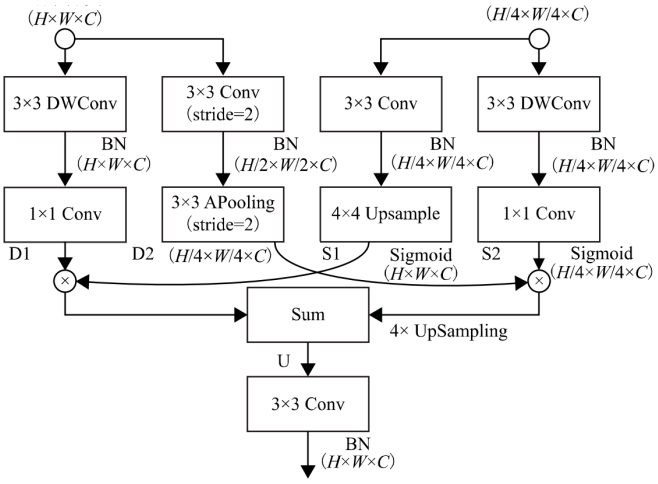

Fig. 9. Aggregation module

$$U = f_2(D_1 \otimes \sigma(S_1), f_1(D_2 \otimes \sigma(S_2) \tag{4}$$

Where, S_1, S_2, D_1 and D_2 are the feature graphs of detail branch and semantic branch processed in the first stage in the aggregation module respectively. U is the fusion feature graph obtained in Sum block. $\otimes$ represents the multiplication of corresponding elements in the feature graph. σ means Sigmoid function. f_1 means to perform 4 times up-sampling operations on the feature graph. f_2 represents the sum of the corresponding elements of the feature graph.

The feature fusion method adopted in this paper pays more attention to the diversity of the upper level road feature and the lower level road feature information in the two branches. By using different scale guidance, different feature representations can be captured, so that multi-scale information can be encoded, so that local information and high-level semantic information can be integrated more effectively, and thus improve the performance of road detection.

4. Experimental results and analysis

The unstructured road detection task is different from the structured road detection task with clear road markers such as expressways and urban arterial roads, which can simplify the detection task into lane lines or road boundary detection without obvious lane lines or clear road boundaries. Traditional urban street view semantic segmentation data sets are not suitable for unstructured road recognition. After comparing with other street view semantic segmentation data sets such as Cityscapes, India Driving Dataset (IDD) is more suitable for the training task of unstructured road detection [36]. The Indian road driving dataset contains 34 categories with a pixel size of 1920×1080. 6792 Street View images that meet the requirements are selected, as shown in Figure 10. As there are many types

of road driving data set in India, targeted transformation is carried out on it. Road and parking are divided into passable areas, and the remaining 32 categories are divided into impassable areas. The selected 6792 images were randomly divided into training set, verification set and test set according to the ratio of 8:1:1. The training set is used to train the network parameters, the verification set is used to feedback the training results and adjust the network parameters in time, and the test set is used to evaluate the quality of the trained network.

(a) (b)

Fig. 10. Selected IDD dataset example

4.1. Network Parameters

The training environment of this network is as follows: Windows11 operating system, CPU Intel(R) Core(TM) i5-10400F, 16GB Memory, Ge Force RTX 3060 GPU, 12GB video memory. We use PyTorch deep learning framework. An adaptive learning rate algorithm named Adam is used to update the weight of the neural network. In training stage, batchsize=8, initial learning rate=1.2×10^{-6}, weight attenuation=5.3×10^{-4}.

4.2. Evaluation Index

In this paper, the Dice similarity coefficient is used as the evaluation index of network model detection accuracy. In order to evaluate the road test results as comprehensively as possible, this paper also uses the recall rate and accuracy rate as evaluation indexes. The calculation formulas of Dice coefficient, Recall and Precision are shown in Equations(5) (7).

$$Dice = \frac{2TP}{FP + 2TP + FN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Where, TP is the number of pixels in the predicted and actual road areas. FP is the number of pixels predicted to be road area but actually non-road area. TN is the number of pixels in the predicted and actual non-road area. FN is the number of pixel points predicted to be non-road area, but actually is road area.

On the other hand, frames per second (FPS) is used as the evaluation index for road detection speed in road area.

$$v = \frac{N}{\sum_i^N t_i} \tag{8}$$

Where N is the number of test images. t is the time required to process the $i - th$ image. v is the number of images processed per second by the semantic segmentation network model.

4.3. Optimized network results comparison

In order to analyze the performance of the network model proposed in this paper and verify the stability of the optimization module, experiments are carried out on the road data set for the road detection task in the open-pit mining area. Four groups of ablation experiments were set up here to compare the performance of BiSeNetV2, BiSeNetV2+SE, BiSeNetV2+FPA and DAM-BiSeNetV2, respectively. The loss function of the training process is shown in Figure 11, and the evaluation index is shown in Table 4.

Table 4. Font

Model	Parameter size/M	Dice/%	Recall/%	Precision/%	FPS/fps
BiSeNetV2	2.34	95.9	93.8	97.9	55
BiSeNetV2+SE	2.37	96.9	96.8	96.8	53
BiSeNetV2+FPA	2.30	97.4	97.3	96.7	69
DAM-BiSeNetV2	2.33	98.5	98.4	97.6	64

According to the analysis of loss function, the four models all reach convergence in the 60 generation times, and the DAM-BiSeNetV2 training loss can be reduced to a smaller loss value compared with other models. Then, the evaluation index of the model is analyzed. In terms of the number of parameters, FPA module adopts deep separable convolution, which reduces the number of parameters compared with ordinary convolution, and also reduces the number of parameters compared with the original CE module by 0.04 M. Combined with SE module, the complexity of the model can be reduced and

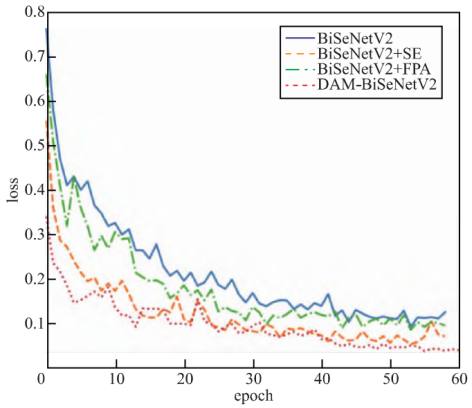

Fig. 11. Comparison of Loss Functions with Different Optimized Networks in Training Set

the reasoning speed of the model can be effectively improved. In terms of accuracy and detection rate, BiSeNetV2 can achieve good results in road detection, but in terms of Dice similarity coefficient and recall rate, it still needs to be further improved. BiSeNetV2+SE benefites from the SE module's ability to extract the dominant relationship between channels. Compared with BiSeNetV2, the Dice coefficient increased by 3.0%, and the tradeoff between recall rate and accuracy is achieved as far as possible. BiSeNetV2+FPA combined with lightweight FPA module can extract features of global concern and reduce model complexity. Compared with BiSeNetV2, the detection rate increases by 14fps and Dice similarity coefficient increases by 1.5%. DAM-BiSeNetV2 integrates SE module and FPA module reasonably. Due to the advantage of introducing multiple feature extraction modules without increasing the number of parameters, the Dice similarity coefficient increases by 2.6%, and the detection frame rate reaches 64fps, which is 9fps higher than BiSeNetV2. When the Dice similarity coefficient is improved, the detection rate of the model for the image is also improved. Therefore, the optimized DAM-BiSeNetV2 model can be better applied to road detection.

4.4. Comparison with other state-of-the-art methods

In order to further analyze the performance of the proposed model for road detection, the proposed method is compared with SSCNN [37], FNN [38], LWIR [39] and CODA [40] on road data sets. To keep the model lightweight, SSCNN and FNN use mobilenetv2 as the backbone, while LWIR and CODA use the backbone from the original network.

The experimental results are shown in Figure 12, where the first two behaviors detect the straight road scene and the last two behaviors detect the curve scene. SSCNN model is unable to detect crooked roads near mountain sides. FNN model will detect the sky or the slope along the road as the road area in some subtle places, and also cannot accurately detect the road area on some curved hillside roads. LWIR has poor detection effect on road images, and many hollowed out areas will appear, which cannot be accurately detected. CODA network model has a good detection effect on the road, but the detection is not

fine enough in the road edge area, and the slope will be mistakenly detected as a road. The method in this paper uses compression excitation module, so the model has the best detection results on the road. From the overall detection effect, the results of the proposed method are better than those of the lightweight LWIR and CODA semantic segmentation networks.

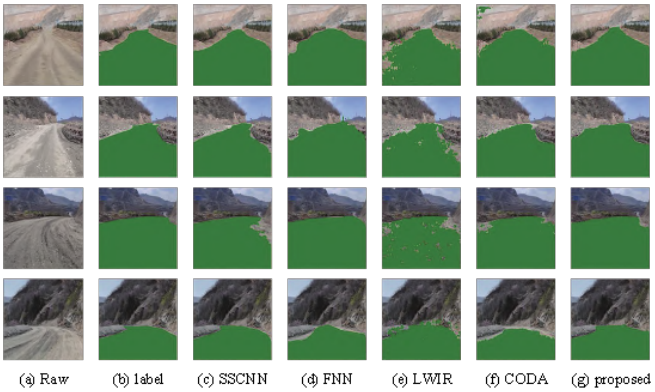

Fig. 12. Results of road detection by different segmentation models

The evaluation indicators obtained are shown in Table 5. The Dice similarity coefficient of the proposed model in this paper reaches 97.5%, and the detection speed reaches 65fps. As SSCNN uses dilated convolution, although the road detection effect is slightly better than that of the proposed model, the detection speed is far lower than that of the new model. Because FNN uses the feature pyramid pool module, it also has a serious impact on the reasoning speed of the model. When compared with LWIR, the proposed method is inferior to this network in terms of detection speed, but LWIR has poor effect on road detection and cannot meet the accuracy requirements. Compared with CODA, proposed method in this paper has 2.5% improvement in Dice similarity coefficient and 9fps improvement in reasoning speed. After the above analysis, the proposed model is superior to other mainstream models in road detection.

Table 5. Road detection results with different segmentation models

Model	Dice/%	Recall/%	Precision/%	FPS/fps
SSCNN	97.8	98.6	96.9	10
FNN	96.0	96.5	95.7	13
LWIR	93.8	91.8	94.4	71
CODA	95.0	93.1	98.9	56
Proposed	97.5	98.4	96.6	65

We also make a comparison on other images as shown in figure 13. By comparing the segmentation effect, it can be seen that the proposed method in this paper is closest to the label image. For the segmentation of the first three images, although the edge segmentation is better in reference [41], it is prone to false segmentation. As shown in the segmentation results of the first and second samples, small pieces of mis-segmentation appears on both vehicles, and the segmentation of the lower edge of vehicles in the segmentation results of the second sample is not very accurate. Although FNN is the fastest, the segmentation accuracy is relatively poor, and the edge details are lost more, which makes it impossible to accurately identify the passable area and the edge of obstacles. LWIR segmentation is more accurate, but the vehicle edge segmentation in the second sample is not accurate, and the segmentation speed is slow. For the fourth sample, except CODA, the other three algorithms can barely segment the left side of the pit contour. But for the right curb, different degrees of mis-segmentation appear in FNN, LWIR and CODA, while the proposed method avoids this phenomenon.

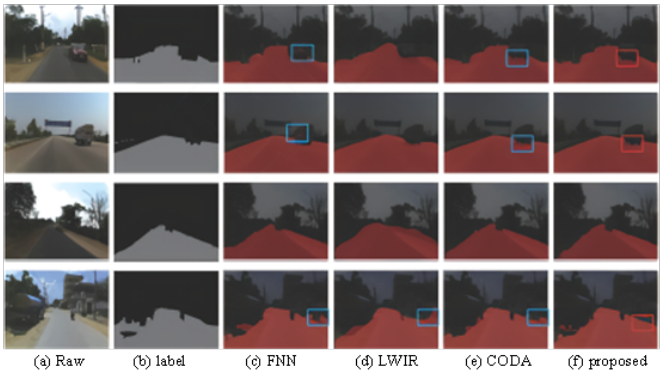

Fig. 13. Comparison of segmentation results

In the same environment and configuration, the selected Indian road driving data set is used to train the network in this paper, and the results are compared with the above networks, as shown in Table 6 (for figure 13). Due to the integration of attention mechanism and lightweight, the new network has improved the performance of average intersection ratio and passable area intersection ratio. The average pixel accuracy is increased to 98.7%, and the frame per second is increased to 72%, reaching 39.4 frames/s, which meets the real-time requirements of unstructured road segmentation task.

5. Conclusions

In order to better fulfill the task of unstructured road recognition, this paper uses selected Indian road driving data sets. Indian road driving data sets are very similar to the unstructured road condition in our country. After training, a semantic segmentation network model suitable for the unstructured road condition can be obtained. Aiming at the problem that structured road segmentation method cannot meet the real time and accuracy

Table 6. Road detection results with different segmentation models

Model	Dice/%	Recall/%	Precision/%	FPS/fps
FNN	96.3	97.4	97.5	32
CLWIR	95.7	92.5	93.8	53
CODA	95.4	94.6	95.2	61
Proposed	98.6	99.1	98.7	72

required by unstructured road segmentation in practical application, this paper proposes a lightweight unstructured network based on BiSeNetV2. To avoid the over-fitting phenomenon caused by too deep network and the low efficiency caused by too many parameters, MobileNetv2 is used to replace Xception as the feature extraction network. In addition, grouping convolution is used to replace common convolution in MobileNetv2 and dilated space pyramid pooling module. Their batch specification layer is rationally selected, which greatly reduces the number of network parameters. This method can improve the segmentation speed without affecting the accuracy of segmentation and meet the requirement of real-time performance. At the same time, in order to enlarge the common features of unstructured road image distribution to improve the segmentation accuracy, BiSeNetV2 is added to output advanced image features. Experiments show that the proposed network is superior to other advanced networks in terms of accuracy and efficiency. It provides some basis for improving the application of semantic segmentation in unstructured road recognition tasks. Future works will focus on more advanced deep learning method for road detection and apply them in real engineering applications.

Acknowledgments. This work was supported by "Project Name: Research on Road Scene Segmentation Technology Based on Convolutional Neural Network; Project Number: 222102210318".

References

1. Teng, L., Qiao, Y. "BiSeNet-oriented context attention model for image semantic segmentation," *Computer Science and Information Systems,* vol. 19, no. 3, pp. 1409-1426. (2022)
2. Liu, D., Yin, S., et al,. "Research on the online parameter identification method of train driving dynamic model," *International Journal of Computational Vision and Robotics*. (2022) doi:10.1504/IJCVR.2022.10047951
3. Rajendar, S., Rathinasamy, D., Pavithrao, R., et al,. "Prediction of stopping distance for autonomous emergency braking using stereo camera pedestrian detection," *Materials Today: Proceedings,* Vol. 51, pp. 1224-1228. (2022)
4. Raipuria. G., Gaisser. F., and Jonker. P.P. "Road infrastructure indicators for trajectory prediction," *2018 IEEE Intelligent Vehicles Symposium (IV). IEEE*, pp. 537-543. (2018)
5. S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan and X. Hu. "Hierarchical Interpretable Imitation Learning for End-to-End Autonomous Driving," *IEEE Transactions on Intelligent Vehicles,* vol. 8, no. 1, pp. 673-683. (2023) doi: 10.1109/TIV.2022.3225340.
6. Liguo Wang, Yin Shoulin, Hashem Alyami, et al. "A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images," *Geoscience Data Journal,* (2022). https://doi.org/10.1002/gdj3.162

7. Kong J, Yang C, Xiao Y, et al. "A graph-related high-order neural network architecture via feature aggregation enhancement for identification application of diseases and pests," *Computational Intelligence and Neuroscience,* vol. 2022, (2022).
8. Abdalla Y, Iqbal M T, Shehata M. "Copy-move forgery detection and localization using a generative adversarial network and convolutional neural-network," *Information,* vol. 10, no. 9, pp. 286 (2019).
9. Zhu J, Yang H, Lin W, et al. "Group re-identification with group context graph neural networks," *IEEE Transactions on Multimedia,* vol. 23, pp. 2614-2626 (2020).
10. Wang H, Wang Y, Zhao X, et al. "Lane detection of curving road for structural highway with straight-curve model on vision," *IEEE Transactions on Vehicular Technology,* vol. 68, no. 6, pp. 5321-5330 (2019).
11. Huang D Y, Chen C H, Chen T Y, et al. "Vehicle detection and inter-vehicle distance estimation using single-lens video camera on urban/suburb roads," *Journal of Visual Communication and Image Representation,* vol. 46, pp. 250-259 (2017).
12. Zhang, J., Yu, X., Lei, X., Wu, C. "A Novel Deep LeNet-5 Convolutional Neural Network Model for Image Recognition," *Computer Science and Information Systems,* Vol. 19, No. 3, pp. 1463-1480. (2022). https://doi.org/10.2298/CSIS220120036Z
13. Li Y, Zhao H, Qi X, et al. "Fully convolutional networks for panoptic segmentation," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 214-223, (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,* vol. 39, no. 12, pp. 2481-2495 (2017).
15. Harika A, Sivanpillai R, Variyar V V S, et al. Extracting Water Bodies in RGB Images Using DEEPLABV3+ Algorithm[J]. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2022, 46: 97-101.
16. Li Y, Li C, Li X, et al. A comprehensive review of Markov random field and conditional random field approaches in pathology image analysis[J]. Archives of Computational Methods in Engineering, 2022, 29(1): 609-639.
17. Yurtkulu S C, ahin Y H, Unal G. Semantic segmentation with extended DeepLabv3 architecture[C]//2019 27th Signal Processing and Communications Applications Conference (SIU). IEEE, 2019: 1-4.
18. Zhou J, Hao M, Zhang D, et al. Fusion PSPnet image segmentation based method for multifocus image fusion[J]. IEEE Photonics Journal, 2019, 11(6): 1-12.
19. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted InterventionCMICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.
20. Yin Shoulin, Liu Jie, Teng Lin. A new krill herd algorithm based on SVM method for road feature extraction[J]. Journal of Information Hiding and Multimedia Signal Processing, vol. 9, no. 4, pp. 997-1005, July 2018.
21. Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
22. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.
23. Guo Q, Wu X J, Kittler J, et al. Self-grouping convolutional neural networks[J]. Neural Networks, 2020, 132: 491-505.
24. Fan J, Hua Q, Li X, et al. Biomedical sensor image segmentation algorithm based on improved fully convolutional network[J]. Measurement, 2022, 197: 111307.
25. Wang W, Wang S, Li Y, et al. Adaptive multi-scale dual attention network for semantic segmentation[J]. Neurocomputing, 2021, 460: 39-49.

26. Chen Z, Rajamanickam L, Tian X, et al. Application of Optimized Convolution Neural Network Model in Mural Segmentation[J]. Applied Computational Intelligence and Soft Computing, 2022, 2022.
27. Kazerouni I A, Dooly G, Toal D. Ghost-UNet: An asymmetric encoder-decoder architecture for semantic segmentation from scratch[J]. IEEE Access, 2021, 9: 97457-97465.
28. Xu Q, Zhang C, Zhang L. Denoising convolutional neural network[C]//2015 IEEE International Conference on Information and Automation. IEEE, 2015: 1184-1187.
29. Zhou D, Wang N, Peng C, et al. Towards multi-domain face synthesis via domain-invariant representations and multi-level feature parts[J]. IEEE Transactions on Multimedia, 2021, 24: 3469-3479.
30. Lin G, Kong L, Liu T, et al. An antagonistic training algorithm for TFT-LCD module mura defect detection[J]. Signal Processing: Image Communication, 2022, 107: 116791.
31. Yin Shoulin, Liu Jie, Li Hang. A Self-Supervised Learning Method for Shadow Detection in Remote Sensing Imagery[J]. 3D Research, vol. 9, no. 4, December 1, 2018. https://doi.org/10.1007/s13319-018-0204-9
32. Li G, Yu Y. Deep contrast learning for salient object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 478-487.
33. Zhu J, Li X, Gao C, et al. Unsupervised community detection in attributed networks based on mutual information maximization[J]. New Journal of Physics, 2021, 23(11): 113016.
34. Liao Y, Lu S, Yang Z, et al. Depthwise grouped convolution for object detection[J]. Machine Vision and Applications, 2021, 32: 1-13.
35. Wang S H, Zhou Q, Yang M, et al. ADVIAN: Alzheimer's disease VGG-inspired attention network based on convolutional block attention module and multiple way data augmentation[J]. Frontiers in Aging Neuroscience, 2021, 13: 687456.
36. Dokania S, Hafez A H, Subramanian A, et al. IDD-3D: Indian Driving Dataset for 3D Unstructured Road Scenes[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 4482-4491.
37. Dewangan D K, Sahu S P. Road detection using semantic segmentation-based convolutional neural network for intelligent vehicle system[C]//Data Engineering and Communication Technology: Proceedings of ICDECT 2020. Springer Singapore, 2021: 629-637.
38. Mehtab S, Yan W Q. Flexible neural network for fast and accurate road scene perception[J]. Multimedia Tools and Applications, 2022, 81(5): 7169-7181.
39. Li N, Zhao Y, Pan Q, et al. Illumination-invariant road detection and tracking using LWIR polarization characteristics[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 180: 357-369.
40. Li K, Chen K, Wang H, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving[C]//Computer VisionCECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23C27, 2022, Proceedings, Part XXXVIII. Cham: Springer Nature Switzerland, 2022: 406-423.
41. Y. Yuan, Z. Xu and G. Lu, "SPEDCCNN: Spatial Pyramid-Oriented Encoder-Decoder Cascade Convolution Neural Network for Crop Disease Leaf Segmentation," in IEEE Access, vol. 9, pp. 14849-14866, 2021, doi: 10.1109/ACCESS.2021.3052769.

Yachao Zhang is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction: Information security, image processing.

Min Zhang is with the Zhengzhou Commercial Technician College. Research direction: Information security, image processing.

Received: December 05, 2022; Accepted: May 10, 2023.