Zhang *et al.*
Chinese Journal of Mechanical Engineering (2024) 37:45
https://doi.org/10.1186/s10033-024-01018-4

Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE

Open Access

Unstructured Road Extraction in UAV Images based on Lightweight Model

Di Zhang[1], Qichao An[1,2], Xiaoxue Feng[1], Ronghua Liu[1,3], Jun Han[1] and Feng Pan[1*]

Abstract

There is no unified planning standard for unstructured roads, and the morphological structures of these roads are complex and varied. It is important to maintain a balance between accuracy and speed for unstructured road extraction models. Unstructured road extraction algorithms based on deep learning have problems such as high model complexity, high computational cost, and the inability to adapt to current edge computing devices. Therefore, it is best to use lightweight network models. Considering the need for lightweight models and the characteristics of unstructured roads with different pattern shapes, such as blocks and strips, a TMB (Triple Multi-Block) feature extraction module is proposed, and the overall structure of the TMBNet network is described. The TMB module was compared with SS-nbt, Non-bottleneck-1D, and other modules via experiments. The feasibility and effectiveness of the TMB module design were proven through experiments and visualizations. The comparison experiment, using multiple convolution kernel categories, proved that the TMB module can improve the segmentation accuracy of the network. The comparison with different semantic segmentation networks demonstrates that the TMBNet network has advantages in terms of unstructured road extraction.

Keywords Unstructured road, Lightweight model, Triple Multi-Block (TMB), Semantic segmentation network

1 Introduction

As the core subject of modern transportation systems, roads are highly significant in terms of geography, politics, and economies. Moreover, roads are also the main components of recording and marking objects in the transportation and global geographic information systems [1]. In the field of modern smart agriculture, the efficient extraction of farmland roads from aerial images helps to quickly divide farmland areas, greatly improving the statistical efficiency of cultivated land. Therefore, the use of modern unmanned aerial vehicle (UAV) photography to extract road information accurately and quickly is crucial to the development of agriculture, and it has gradually become a research hotspot.

Rural farmland roads are representative examples of unstructured roads. There are no unified planning indicators for roads, and the roads have complex and variable morphological structures. Owing to the complex shape features and blurred road edges, road extraction becomes more difficult. Traditional road extraction algorithms based on different prominent features of roads in images include threshold segmentation algorithms based on image pixel statistics [2, 3], edge segmentation algorithms based on image road edge details [4], extraction algorithms based on road geometric features [5], and road extraction algorithms based on the probability graph model [6–8]. However, these algorithms exhibit poor performance in road extraction.

In road segmentation algorithms based on deep learning, numerous network parameters are used

*Correspondence:
Feng Pan
panfeng@bit.edu.cn
[1] School of Automation, Beijing Institute of Technology, Beijing 100081, China
[2] Tianjin Institute of Maritime Navigation Instruments, Tianjin 300130, China
[3] The 716th Research Institute of China Shipbuilding Group Co., Ltd, Jiangsu 222061, China

to learn feature information from images. In road extraction, road segmentation can be regarded as a semantic segmentation problem [9]. Many studies on road extraction based on deep learning have demonstrated effective extraction of structured roads [10, 11]. Li et al. [12] and others proposed a linear integrated convolution algorithm using a large sample dataset based on a convolutional neural network. By predicting the probability that a pixel region in an image is part of a road, this approach determines whether each image pixel belongs to a road area. Xin et al. [13] used an improved UNet algorithm to improve the accuracy of the network for road segmentation through dense links and skip connections.

Current road extraction algorithms based on deep learning technology mostly focus on improving the accuracy of road segmentation algorithms, while ignoring the forward inference speed. However, this approach has the disadvantages of high model complexity and high computational cost, making it unable to adapt to current edge computing devices. Balancing the accuracy and real-time performance of a network model and designing a lightweight unstructured road extraction model have significant research value. This study attempted to design a lightweight network model for unstructured road extraction.

In the traditional encoding-decoding network structure, the shallow network in the encoder contains detailed features, such as edges, which are helpful for edge segmentation. Deep networks can obtain large receptive fields from small-resolution input images, thereby capturing global semantic information. The semantic information in the segmentation object is extracted by the deep network, which provides the basis for determining the category of the segmentation object. For feature information fusion, in the encoding stage, the equal-proportional fusion method is used; this approach does not consider the impact of the feature information contained in different feature layers for the final segmentation result when the encoder encodes the information. A large amount of information is redundant, and this method has the problems of road breaks and blurred edges in the results of unstructured road segmentation. In the design of a lightweight semantic segmentation network model, methods such as separable convolution, convolution kernel decomposition, and dilated convolution are used to reduce the number of parameters and calculations of the network model, reduce the downsampling rate, and output larger feature maps, retaining more spatial information. Bilinear interpolation does not require additional parameters or backpropagation calculations, can run quickly, and is widely used in the upsampling design of lightweight models for semantic segmentation [14–16]. The lightweight road extraction network model designed in this study uses bilinear interpolation to restore the feature sizes through upsampling.

In convolutional neural networks, only a small number of parameters are required to achieve accurate prediction results [17], which shows that there is a large amount of information redundancy in the process of convolution operations on images performed by the convolution kernel. Therefore, improving the convolution kernel can reduce the number of parameters and information redundancy. Channelwise convolution is often combined with point convolution to achieve depthwise separable convolution [18, 19]. MobileNet [20–22] combines channel-by-channel convolution and point convolution to form a separable convolution method that reduces the number of calculations performed by the convolution kernel. Inception [23] uses 1×1 convolution to reduce the feature dimension, and uses the superposition of different convolution kernels to fuse information between high- and low-level features under the same feature layer. Simultaneously, multi-scale receptive fields are used to obtain contextual information in different ranges to improve the accuracy of the network. ShuffleNet [24] uses grouped convolution for convolution, cross-mixes the feature information between different groups, and reduces the number of calculations through grouped convolution. The fusion of feature information between different groups can make full use of the exchange of information to improve network performance. The above network uses changes in the convolution method, as well as grouped convolution, decomposable convolution, etc., to reduce the number of calculations and parameters. ENet [25] adopts the design concept of ResNet [26] and is designed using a lightweight bottleneck module. This module uses 1×1 convolution to reduce the number of input channels. After extracting features through the middle convolution module, the number of features is restored through 1×1 convolution, which significantly reduces the number of model parameters. Additionally, the diversified convolution kernel construction method enhances the model's ability to extract different features. The convolution decomposition design concept of ENet provided ideas for the network structure design of ERFNet [27], ESPNet [28], and other networks to a certain extent. However, ENet focuses excessively on efficiency improvement. Although it reduces the number of model parameters, the model accuracy also decreases significantly.

ERFNet takes advantage of decomposition convolution and proposes a one-dimensional non-bottleneck module to reduce the number of calculations. Additionally, it uses dilated convolution in the one-dimensional

non-bottleneck module to capture more contextual information and enter the next layer of the network. Excessive downsampling of feature maps results in the loss of spatial information of road targets and increases the number of parameters and calculations of the network, resulting in a further loss of accuracy and model forward inference speed in road extraction [29]. The purpose of dilated convolution is to expand the range of the receptive field and retain the spatial information of the road targets in the image without downsampling the feature map. Although EDANet [14] maintains the use of decomposed convolutions to form the basic asymmetric convolution module, it also introduces the concept of dense connections, which can jointly collect multi-scale information, allowing EDANet to achieve good segmentation performance at a low computational cost. CGNet [30] introduces a context module in the feature extraction process of the encoder, using joint learning of the local features of the target and contextual information around the target to improve network segmentation accuracy. LEDNet makes full use of decomposed convolution and grouped convolution to fully integrate the features between different channels while reducing the computational complexity of the network.

Compared with traditional network models that pursue accuracy, such as DeepLab [31], lightweight network models must make fuller use of the advantages of decomposed convolution, separable convolution, and atrous convolution in the structure design. While ensuring that the encoder can fully extract the image feature information, the number of parameters and network calculations must be reduced, and a balance must be achieved between the accuracy and inference speed of the model. In the lightweight road extraction network introduced in this paper, the application of depth-separable convolution and atrous convolution can reduce the amount of feature network parameters, expand the scope of the receptive field, retain more spatial information of road targets, and extract information features of road targets in images.

A multi-branch semantic segmentation model uses different branches to extract different features from an input image. ICNet adopts the structural form of PSPNet [32] and uses three different branches to extract features from input images at different resolutions. For feature fusion, high-order contextual information is used to perform layer-by-layer upsampling and combine the result with high-order features. The refined fusion method fully utilizes the feature information extracted from different branches to obtain the final segmentation result. ContextNet [33] uses a dual-branch structure network, taking images of different resolutions as inputs and applying separable convolution to further reduce the number of network parameters and calculations. Fast-SCNN [34] aims to solve the problem in which the initial convolution features cannot be shared in the dual-branch network. It performs initial convolution on the input image to obtain the feature map and uses the dual-branch network to extract subsequent high-order semantic features and spatial detail information features from the feature map. ContextNet and Fast-SCNN adopt dual-branch network structures to segment different targets based on lightweight models.

BiseNet [35] adopts a multi-branch design structure and uses a fast downsampling context path to extract context information. Spatial paths are used to produce high-resolution output features to preserve the spatial information. After feature fusion of the feature maps generated by the two branches, the final segmentation result is output through upsampling. ShelfNet [36] provides a more complex network structure design, using a multi-path convolutional neural network, and draws on the design ideas of recurrent neural networks to achieve weight sharing. To a certain extent, ShelfNet can be regarded as a variant of the FCN network structure; however, its network structure is lighter and smaller, and its network inference speed is faster. The lightweight network model developed in this study employs a multi-path approach to extract roads.

The design of a lightweight model structure requires full simplification of the network structure. In the encoder stage, the structures of different convolution kernels are fully utilized to extract multi-scale contextual information. To avoid the loss of spatial information caused by excessive downsampling and increase the receptive field of the convolution kernel, dilated convolution can be added in the network encoder stage. In the decoder stage, linear upsampling is used to restore the size to ensure a lightweight network design.

In this study, we investigated a lightweight semantic segmentation algorithm for unstructured rural farmland roads, adopted a coding-decoding structure design form, and proposed the TMBNet network model. In the encoder stage, the triple multi-block (TMB) module is used to fully extract various contextual information of the input features using multiple paths, multiple receptive fields, and multi-convolution kernels. In addition, the network parameters and number of calculations are significantly reduced. The TMB module has a highly flexible structural form, and can dynamically adjust the number of branches and change the convolution kernel form of the branches simultaneously. In the decoder stage, linear interpolation is used for upsampling to restore the feature map size. The remainder of this paper is organized as follows. Section 2 introduces the

principles of the triple multi-block module. Section 3 describes the architecture of the TMBNet network. Section 4 describes the comparative experiments. Finally, the conclusions are presented in Section 5.

2 Triple Multi-Block Design

For the triple multi-block module, based on the design idea of InceptionNet and on the basis of multi-path and multi-receptive-field input feature extraction, deep separable convolution and asymmetric convolution are used to reduce the number of parameters of the module. A good balance is achieved between input feature extraction, reducing the number of parameters, and reducing the computational cost.

2.1 Multi-path-based Design

The multi-path design method is feasible for implementing different types of convolution kernels in the same module. In terms of structural design, the multi-branch design has a high degree of flexibility, and the number of branches can be flexibly increased or decreased. For the multi-path design method, to reduce the number of parameters and the number of calculations, 3×3 convolution is used to extract the input features while reducing the channel dimension of the input features. Then, different branches are used to perform feature extraction on the same feature, and the features are superimposed according to the feature channel dimension. The corresponding mathematical expression is given in Eq. (1):

$$Map_{out} = \underset{i=1}{\overset{N}{Concat}} f\left(\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} k_i \times x_{h,w} + b\right), \tag{1}$$

where N represents the number of paths; k_i represents the convolution kernel types of different branches; C, H, W respectively represent the number of channels inputting different branch features and the height and width of the feature map; and f represents the activation function.

2.2 Design Analysis Based on Multiple Convolution Kernels

Because there is no unified planning and design method for unstructured roads, there are large differences in the shapes and structures of roads, and different modes, such as block and strip shapes, exist. The multi-convolution kernel design mode can extract specific mode response features for unstructured roads with different structures. In Figure 1, the schematic on the left shows an enlarged, detailed display of the convolution on the right. On the left side of the figure, the yellow module represents the convolution kernel, and the black and white areas represent the background and road, respectively. For the block road area, feature information can be quickly extracted by using the convolution kernel of the N×N structure. For the feature information of long and narrow roads, the asymmetric structure of the $N \times 1$ convolution kernel is used to reduce the number of parameters and calculations, and it provides a good feature response to long and narrow roads. The asymmetric convolution kernel performs convolution calculations along the edge of the road. Compared with the N×N symmetric structure convolution kernel, the asymmetric convolution kernel is better able to distinguish between road and non-road information under a slender road structure. Different types of convolution kernels can be used to respond to specific patterns of road structures with different pattern characteristics, thereby improving the network's road segmentation accuracy.

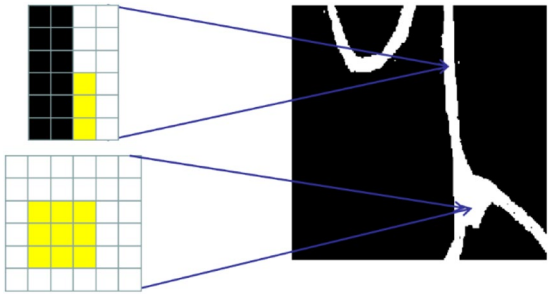

Figure 1 Convolution diagram of multiple convolution kernels

2.3 Design Analysis Based on Multiple Receptive Fields

For road extraction, the road in the image used as saliency information exhibits significant variation in size and spatial position information. The single structure of the convolution kernel limits the range of the receptive field, and the range of the receptive field in practical applications is smaller than the theoretical receptive field [37]. The contextual information in the input features cannot be fully mined. Thus, this approach is significantly limited in its ability to capture the characteristic information of roads of different sizes. A wide range of receptive fields is conducive to the acquisition of global information, whereas a small range of receptive fields is advantageous for the extraction of local information. Therefore, by using multiple convolution kernels, feature information under different receptive fields can be obtained in the same module, and global and local information can be better captured. The size of the receptive field of the convolution kernel in the feature map is calculated using Eq. (2):

$$RF_{l+1} = RF_l + (k_{l+1} - 1) \times s_l. \tag{2}$$

For the dilated convolution, the receptive field size is calculated using Eq. (3):

$$RF_{l+1} = RF_l + (k_{l+1} - 1) \times s_l \times d_{l+1}, \tag{3}$$

where RF represents the size of the receptive field, l represents the number of layers, s represents the step length information, and d represents the dilation rate.

2.4 Triple Multi-Block Module

The structure of the TMB module is shown in Figure 2. In the TMB module, 3×3 convolution is used for input feature extraction and the number of channels for the output features is reduced according to the number of branches, as shown in Eq. (4), where N_{path} denotes the number of branches:

$$Channel_{out} = Channel_{in}//N_{path}. \tag{4}$$

According to Eq. (4), the number of output channels is reduced based on changes in N_{path}, so there will not be a significant increase in the number of calculations or the number of parameters. In different branches, all convolution operations adopt the operation mode of depthwise separable convolution, which reduces the number of calculations and the number of parameters, while performing regularization to prevent the network from overfitting. Each branch makes full use of the key ideas of lightweight model design, and uses factorized convolution, dilated convolution and other methods to form receptive fields of different sizes. After the features are superimposed, the 1×1 convolution operation is used to fuse the feature information under different receptive fields to further enhance the exchange of global information and local information and to improve the understanding of unstructured road information. Using the idea of residual networks as a reference, the input feature and output feature are added together, and the result is output to the next module.

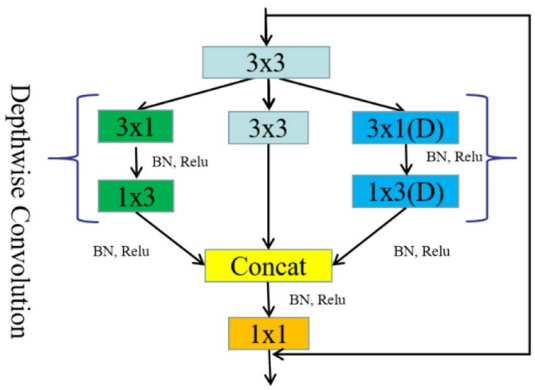

Figure 2 TMB module

In the TMB module shown in Figure 2, the input feature map passes through a 3 × 3 convolution and then enters three branches. On the left, 3 × 1 and 1 × 3 convolution kernels are successively applied. The middle branch uses a 3 × 3 convolution kernel. On the right side, 3 × 1 and 1 × 3 convolution kernels are successively applied, and dilated convolution is used. After the fusion of the three branches, a 1 × 1 point convolution is used to change the number of output channels, and the result is finally fused with the input feature map.

In Figure 2, assuming that the initial receptive field size is 1, the stride is 1, and the dilation rate is 2, the receptive field sizes of the different branches are calculated using Eq. (5):

$$\begin{aligned} RF_{left1} &= 1 + (3-1) \times 1 = 3, 1 + (1-1) \times 1 = 1, \\ RF_{middle} &= 1 + (3-1) \times 1 = 3, 1 + (3-1) \times 1 = 3, \\ RF_{right1} &= 1 + (3-1) \times 1 \times 2 = 5, 1 + (1-1) \times 1 \times 2 = 1. \end{aligned} \tag{5}$$

The multi-convolution kernel design method of the TMB module exhibits a characteristic form of multiple receptive fields. Compared with EDANet and other networks, TMBNet has a multi-branch structure and provides path support for multiple receptive fields. Compared with LEDNet, TMBNet adds a multi-convolution kernel structure based on multiple paths and provides a multi-receptive-field feature extraction method.

3 TMBNet Network Structure

Based on the TMB module, the detailed design of the structure of the final TMBNet network model is provided.

TMBNet adopts the encoding-decoding structure of the multi-scale FCN and extracts image features by downsampling the convolutional network in the encoding stage. To improve the inference speed of the network model in the decoding stage, TMBNet uses bilinear interpolation to replace deconvolution with feature upsampling. To fully integrate spatial and semantic information in the downsampling process, TMBNet merges more feature information elements by superimposing feature maps before upsampling to improve the accuracy of the network model. The TMBNet model structure was designed to consider both speed and accuracy, as well as the balance between the forward inference speed and accuracy of the network model. The model structure is shown in Figure 3.

In the encoder part, while the network performs feature extraction through downsampling, a rapid decline in the resolution of the feature map will cause a large loss of spatial information of the road object. Therefore, compared with UNet, FCN, and other networks, the

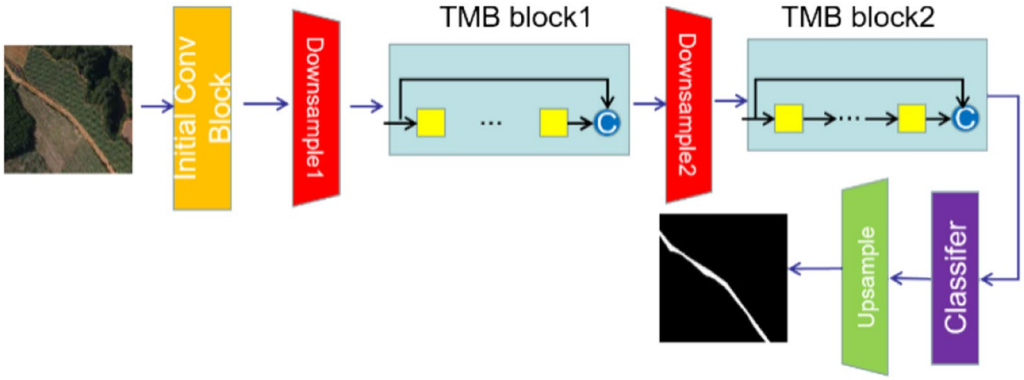

Figure 3 The structure of the TMBNet network

TMBNet network performs downsampling only three times, and the original input image is downsampled eight times. A feature layer with a higher resolution can better retain the spatial information of a segmented object and provide more accurate edge detail information for feature information fusion. In the decoder stage, TMBNet uses an upsampling method of bilinear interpolation to restore the size of the feature map to that of the input image.

The TMBNet network uses the initial convolution module of the three-layer convolutional network to extract the bottom-layer feature information of the image, and the input image is downsampled twice, reducing the number of calculations performed in the subsequent modules. In the downsampling module (Downsample1, Downsample2), the network uses convolution + pooling to halve the feature map in the two dimensions of height and width, and double the channel dimension; then, spatial information is used to exchange feature dimension information. In the TMB Block1 and TMB Block2 components constructed by the TMB module, the input feature map is extracted through the TMB module, and the input and output features are superimposed to fuse the spatial and semantic information of the local TMB component. The TMB component expands the receptive field of the convolution kernel via dilated convolution. By fully capturing the input feature map information, the spatial resolution of the input feature map is retained, and the feature map is not downsampled. The feature map extracted by the TMB Block2 component is passed through a 1×1 convolutional network layer (Classifier) to reduce the dimensionality of the output feature map; the number of channels of the output feature map is the number of classified categories. The final output characteristic result is obtained through bilinear interpolation.

The size of the input network RGB image was set as $256 \times 256 \times 3$ pixels. For the network structure in Figure 3, specific information on the implementation details of the network operation, output feature map size, convolution step size, and dilation rate is given in Table 1.

Table 1 The implementation details of TMBNet

Network layer	operation	Convolution stride/the dilation rate	Output feature channel	Output feature size
1	3×3 Conv	2/1	32	128 × 128
2	3×3 Conv	1/1	32	128 × 128
3	3×3 Conv	1/1	32	128 × 128
4	Downsample	–	64	64 × 64
5–7	TMB	1/2	128	64 × 64
8	Downsample	–	128	32 × 32
9–10	TMB	1/4	128	32 × 32
11–12	TMB	1/8	128	32 × 32
13–14	TMB	1/16	256	32×32
15	1×1 Conv	1/1	2	32 ×32
16	Upsample	–	2	256×256

4 Experimental Results and Discussion

4.1 Network Model Evaluation Criteria Based on Lightweight Road Extraction

To ensure the accuracy and real-time performance of the road extraction network model, as well as the usability of the segmentation network model in practical applications, it is necessary to conduct a comprehensive and fair analysis and comparison between the road extraction network model and existing network models in terms of model segmentation accuracy, forward reasoning speed, and model size. In terms of network model accuracy, the evaluation criteria included the mean intersection over union, accuracy, and recall. To examine the reasoning speed of the network model, the evaluation criteria included Hz (FPS) and the number of model parameters.

(1) Mean Intersection over Union

The intersection over union (IOU) measures the ratio of the number of true positives (intersection) to the sum of true positives, false negatives, and false positives (union). This is an important criterion for measuring the accuracy of image segmentation. As shown in Figure 4, the intersection of the true value (TV) and predicted value (PV) is the true positive (TP), and the union of TV and PV is the true negative (TN). The intersection between TV and non-PV is the false negative (FN) and the intersection between PV and non-TV is the false positive (FP). The IOU is expressed by Eq. (6):

$$IOU = \frac{TP}{TP + FP + FN}. \tag{6}$$

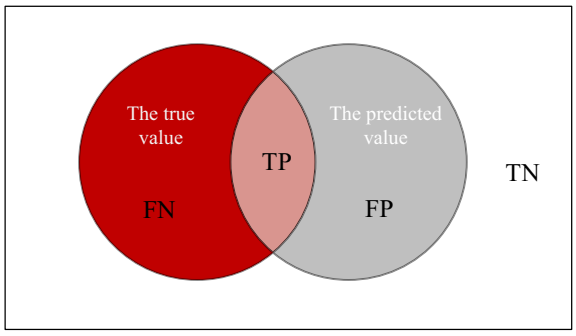

Figure 4 The diagram of intersection over union

For two-dimensional image data, we assumed a total of $k+1$ classes in the image to be classified. p_{ij} denotes a case in which the true class is i, but it is incorrectly classified as class j; this are usually interpreted as FP, whereas p_{ji} represents FN and p_{ii} represents TP. The mIOU can be obtained by calculating the mean intersection over union of the $k+1$ classes, as shown in Eq. (7):

$$mIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}. \tag{7}$$

(2) Accuracy

The accuracy represents the percentage of samples that were correctly predicted relative to the total number of samples, as shown in Eq. (8):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

(3) Recall

The recall represents the ratio of the samples whose value is the true value in the predicted value to all samples with the true value, as shown in Eq. (9):

$$recall = \frac{TP}{TP + FN}. \tag{9}$$

(4) Hz

Hz expresses the number of frames that the model can process in one second, that is, the number of road images that can be extracted per second. The Hz value of the road extraction network model is affected by the computing power of the hardware platform and the size of the input image. Under the same hardware computing conditions and input image size, the larger the Hz number of the model, the faster the forward inference speed.

(5) Number of Model Parameters

The number of parameters of the road extraction model determines the amount of memory required. The traditional road segmentation network necessitates a large number of parameters and high computing power requirements, and cannot

be deployed and operated on mobile terminals. The lightweight road segmentation model can be deployed in edge computing equipment to realize real-time model operation with few parameters and calculations.

4.2 Model Training Details

To train the model, weighted cross-entropy was used as the loss function, and the stochastic gradient descent algorithm was used to optimize the loss function. To adjust the learning rate, the warm-up learning rate adjustment strategy [26] was adopted. In the initial training stage, a smaller learning rate was selected, and when the model iteration reached a preset standard, the learning rate of the model was switched to a preset value. Using the warm-up learning strategy can avoid overfitting in the initial stage of model training while simultaneously improving the deep stability of the model. The decay mode of the learning rate is shown in Figure 5.

The FROBIT training set was used as the training data. In the model training process, the batch size was set to 8, the data were randomly inverted, and random noise was added to enhance the data. The FROBIT test set was used as the test data to evaluate the performance of the model. In terms of model accuracy, the mean intersection over union, accuracy, and recall were used as the evaluation criteria. In terms of the forward reasoning speed, the FPS, model parameters, and storage space required by the model were used as the evaluation criteria.

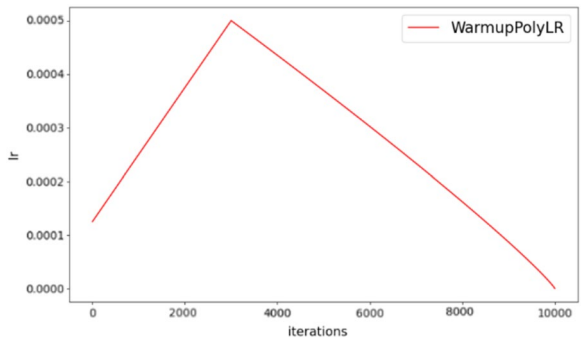

Figure 5 The decay of warm-up learning rate

4.3 Comparative Experiment Based on TMB Module

Based on the control variables, the rationality and effectiveness of the TMB module for unstructured road extraction were observed through ablation comparison experiments between different modules and analysis of the experimental results. Three sets of comparative experiments were conducted using the TMB module.

4.3.1 Experimental Analysis of the Effectiveness of the TMB Module

For the TMBNet network structure, a TMB Block component is constructed using the TMB module. To verify the effectiveness of the TMB module in the overall network architecture, the TMB module was replaced with the feature extraction modules of other networks, and the following four groups of comparison networks were designed.

The Net1 network references the non-dense module in EDANet. This module was used to replace the TMB module in Figure 6. The number of modules was 3 and 6, respectively. The network structure details, excluding the replaced non-dense module, are listed in Table 1. The number of output feature channels changed to 284 in layers 5 to 7 and to 736 in layers 13 to 14. The output feature size remained unchanged.

The Net2 network references the Non-bottleneck-1D module in ERFNet. The Non-bottleneck-1D module was used to replace the TMB module in Figure 6. The number of Non-bottleneck-1D modules was three and six, respectively. Except for the replaced non-bottleneck-1D,

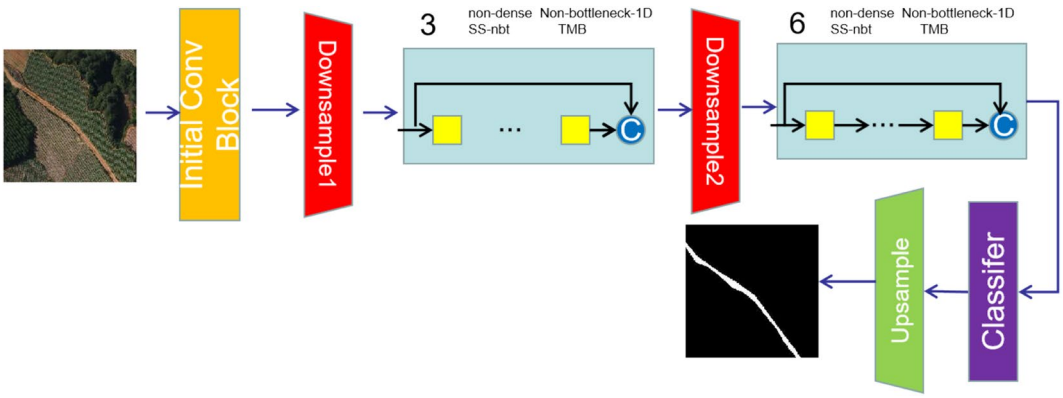

Figure 6 Diagram of test model to evaluate TMB module effectiveness

the network structure details were consistent with those of TMBNet in Table 1.

The Net3 network references the SS-nbt module in LEDNet. The SS-nbt module was used to replace the TMB module in Figure 6. The number of SS-nbt modules was three and six, respectively. Except for the replaced SS-nbt module, the network structure details were consistent with those of TMBNet in Table 1.

The structural diagram of the Net4 network is also shown in Figure 6, containing 3 and 6 TMB modules respectively. The network structure details are listed in Table 1.

From the evaluation indicators in Table 2, it can be seen that Net4 with the TMB module as the core module had the smallest number of network parameters in terms of model computing resources. The number of parameters of Net4 was only 36% of that of Net2. The Channel Shuffle operation of Net3 requires many pointer jumps, and additional feature storage space is required during the calculation process.

During the downsampling process of Net4, there is no need to consume additional storage space or perform complex calculations; thus, the training time of the model is reduced.

Net4 also achieved the best performance in precision comparison indicators, such as the mean intersection over union and recall rate. For example, Net4 has a mean intersection over union 3.57% higher than that of Net2 and 4% higher than that of Net3. The experimental results show that the multi-branch and multi-receptive-field design of the TMB module can fully extract multi-scale feature object information and improve the segmentation accuracy of the network model by ensuring a favorable forward reasoning speed.

To demonstrate the effectiveness of the TMB module in extracting unstructured roads more intuitively, the results of road extraction are visually displayed. A comparison and analysis of the road extraction results for the above four groups of networks on the FROBIT test set are shown in Figure 7. The first and second columns in Figure 7 show the real images in the test set and the unstructured road labels in the images, respectively. The third to sixth columns represent the road extraction results of the real images by different networks. As shown in Figure 7, the three groups of networks from Net1 to Net3 all had road extraction fractures and incomplete road segmentation during the extraction of roads. The continuity and accuracy of Net4 are better than those of the other three groups of comparative networks.

In the road extraction shown in the second row of Figure 7, the surrounding building features of the road in the lower left corner are complex, and there is a

Table 2 Performance comparison of different modules

Network	Parameters	mIOU (%)	Recall	Accuracy	FPS
Net1	0.86 M	77.30	0.7993	0.9515	110
Net2	1.51 M	74.16	0.7655	0.9615	150
Net3	0.85 M	73.73	0.7624	0.9583	108
Net4	0.55 M	77.73	0.8035	0.9521	120

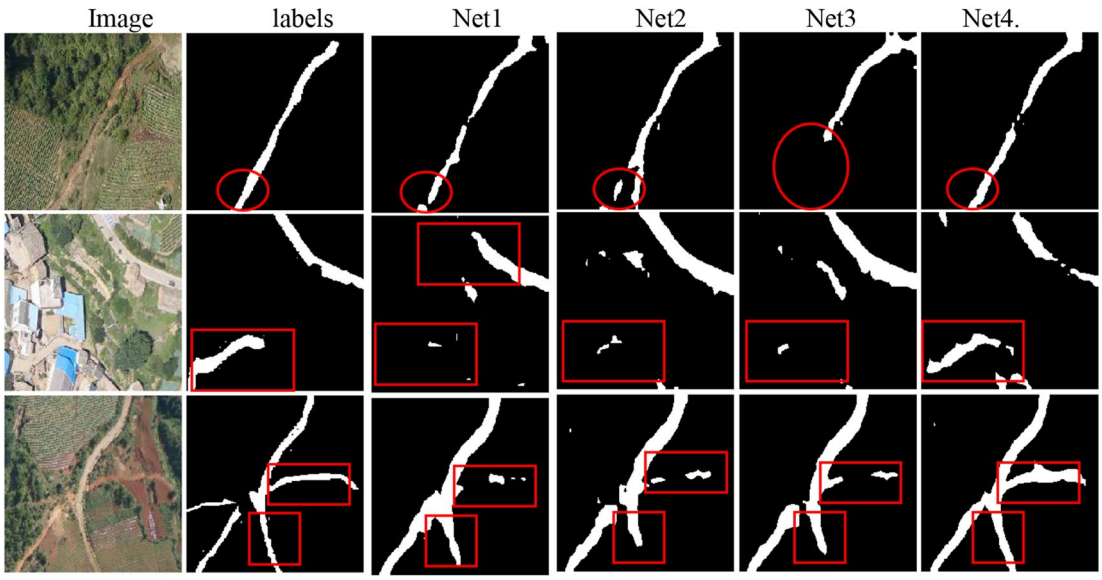

Figure 7 Visual display of FROBIT test set under different network modules

high degree of similarity in color and texture features between the road and the surrounding buildings. Net4 can make better use of the contextual feature information in the image so that the road can be segmented and extracted accurately. The third row presents unstructured roads of different widths. Net4 could extract roads more continuously and completely, reflecting the superior performance of the TMB module as well as its ability to improve the segmentation accuracy of the network.

4.3.2 Experimental Analysis of Multi-convolution Kernel Design

In the experiment on the effectiveness of the TMB module, the effectiveness of the TMB structure was proven by comparing different feature extraction modules. In the experiments described in this section, by comparing the different receptive field sizes under different convolution kernel structures, it was further verified that the convolution method of multiple convolution kernels in the TMB module produces multiple receptive fields, which are conducive to feature extraction. In this experiment, four groups of comparison networks were designed based on the TMB module. The convolution kernel shape of each branch was changed to maintain the three different branches of the TMB module. The method for changing the convolution kernel in the network structure is as follows.

Net_Left network: The left branch of the TMB module was kept unchanged, and the middle and right branches were replaced with the convolution kernel form of the left branch.

Net_Middle network: The middle branch of the TMB module was kept unchanged, and the left and right branches were replaced with the convolution kernel form of the middle branch.

Net_Right network: The right branch of the TMB module was kept unchanged, and the left and middle branches were replaced with the convolution kernel form of the right branch.

TMBNet network: Based on the TMB module, it maintains the convolution form of multiple convolution kernels and receptive fields.

The experimental results are listed in Table 3. The comparative analysis of experiments indicated that the number of parameters of the models was the same in each of the four compared networks. In the process of convolutional feature extraction, each branch of the Net_Middle network performs the convolution process only once; therefore, it has an advantage in terms of FPS. In the comparison of model accuracy, the performance of the network with a single receptive field was similar in terms of the mean intersection over union, recall, and accuracy. The TMBNet network based on multiple receptive fields was better than the other three groups of networks in terms of accuracy, which reflects the advantages of the TMB module. This indicates that the TMB module based on multiple convolution kernels can effectively use receptive fields at different scales and fully extract the object feature information from the input features, thereby improving the segmentation accuracy and performance of the network.

4.3.3 Comparative Experiment Based on the Number of TMB Modules

Based on the comparative experiments described in Sections 4.3.1 and 4.3.2, the effectiveness of the structural design of the TMB module for the extraction of unstructured road features has been proven. In this section, to further verify the influence of the number of TMB modules on the performance of the TMBNet network, in this section, experiments related to the number of TMB modules are reported.

Five groups of comparison networks were designed according to different numbers of TMB modules in TMB block1 and TMB block2 in the TMBNet network. The networks were named following the structure of TMBNet_{TMB block1}_{TMB block2}. For example, TMBNet_3_6 indicates a network in which TMB block1 contains three sets of TMB modules and TMB block2 contains six sets of TMB modules. To ensure that the feature map was not downsampled while expanding the receptive field, the dilated convolution rate of TMB block2 was set as [4, 4, 8, 8, 16, 16, 31]. The experimental results are listed in Table 4. By analyzing the experimental data in Table 4, it can be concluded that the number of parameters and amount of memory required by the

Table 3 Multi-convolution kernel design comparison experiment

Network	Parameters	Memory	mIOU(%)	Recall	Accuracy	FPS
Net_Left	0.55 M	2.35 M	77.21	0.8013	0. 9439	105
Net_Middle	0.55 M	2.17 M	77.07	0.7985	0. 9475	132
Net_Right	0.55 M	2.35 M	77.03	0. 7966	0.9514	105
TMBNet	0.55 M	2.27 M	77.73	0.8035	0.9521	120

network model increase as the number of TMB modules in block1 and block2 increases. In the encoder stage, as the number of network layers increases, the number of output feature map channels increases. For example, TMBNet_3_7 and TMBNet_4_6 contained the same number of TMB modules. However, because the number of input feature map channels in block2 is twice that in block1, TMBNet_3_7 has a larger number of parameters and requires a larger amount of memory storage. TMBNet_2_5 exhibited the best performance in terms of forward reasoning speed and amount of memory required; however, it performed poorly in terms of model accuracy. A comprehensive comparison of the performance of the network model in Table 4 in terms of accuracy and speed shows that TMBNet_3_6 achieved the most reasonable balance between segmentation accuracy and forward reasoning speed.

In response to the poor accuracy of TMBNet_2_5, we added an attention mechanism. The attention mechanism has achieved satisfactory results in image processing. The channel attention mechanism SE-Net (squeeze-and-excitation network) models the interdependence between channels and adaptively determines the importance of each channel [38]. Its core structure consists of two parts: compression and excitation. ECANet improves the excitation module of SE-Net and uses a local cross-channel method to obtain adjacent channel information weights [39]. In this study, based on TMBNet_2_5, an efficient channel attention (ECA) module was added, and a simple experiment was performed. The network had 0.49M parameters, and the mIOU and recall were better than those before adding the ECA module, at 0.7715 and 0.8020, respectively, whereas the accuracy was 0.9404, which is not as good as that before adding the ECA module.

4.4 Semantic Segmentation Network Model Comparison Experiment

To verify the effectiveness of the TMBNet network for unstructured road extraction, it was compared with other semantic segmentation networks in terms of model accuracy and inference speed. The experiment was divided into two parts: (1) Comparing TMBNet with traditional networks in terms of accuracy improvement, and (2) comparing TMBNet with other lightweight semantic segmentation models. The experimental results are

Table 4 Comparison network based on the number of TMB modules

Network	Parameters	Memory	mIOU(%)	Recall	Accuracy	FPS
TMBNet_3_6	0.55 M	2.27 M	77.73	0.8035	0.9521	120
TMBNet_3_7	0.61 M	2.50 M	76.90	0.8012	0.9361	108
TMBNet_4_6	0.57 M	2.34 M	77.52	0.8011	0.9528	105
TMBNet_2_5	0.49 M	1.96 M	77.01	0.7997	0.9429	130
TMBNet_4_7	0.63 M	2.56 M	77.79	0.8072	0.9444	93

Table 5 Comparing TMBNet and lightweight semantic segmentation network

Network	Parameters	Training time	mIOU(%)	Recall	Accuracy	FPS
CGNet	0.49 M	12 h 46 min	77.67	0.8043	0.9483	80
Liteseg	4.38 M	7 h 39 min	76.41	0.7933	0.9433	110
FDDWNet	0.81 M	24 h 30 min	72.59	0.7551	0.9431	65
LEDNet	0.92 M	56 h 36 min	75.10	0.7776	0.9507	86
TMBNet	0.55 M	6 h 7 min	77.73	0.8035	0.9521	120

Table 6 Comparing TMBNet and accuracy improvement semantic segmentation network

Network	Parameters	Training time	mIOU(%)	Recall	Accuracy	FPS
PSPNet	53.58 M	92 h 40 min	78.94	0.8136	0.9577	39
LinkNet	11.53 M	8 h 19 min	79.40	0.8229	0.9473	96
UNet	7.78 M	40 h 12 min	80.57	0.8364	0.9454	95
TMBNet	0.55 M	6 h 7 min	77.73	0.8035	0.9521	120

presented in Tables 5 and 6. From the data in Tables 5 and 6 in the comparison of lightweight networks, CGNet has advantages in terms of the number of parameters and can achieve better accuracy; however, a long training time is required and the model's forward reasoning speed is slow.

In the comparison of mIOU, TMBNet achieved the best performance, achieving a 1.32% improvement compared to Liteseg. At the same time, the forward inference speed of TMBNet is approximately twice that of FDDWNet.

Compared with the network with improved accuracy, TMBNet can achieve significant advantages in terms of model parameters and forward reasoning speed with less accuracy loss. Compared with the PSPNet network, the number of parameters of TMBNet was only 1.03% of that of PSPNet, and the accuracy rate was only reduced by 1.21%. Through comparative experiments with different network models, it was demonstrated that the TMBNet can achieve a more efficient forward reasoning speed using fewer parameters and computational resources when extracting unstructured roads, and it can provide better results for this task.

5 Conclusions

This article first discusses the necessity of lightweight network design and proposes the TMB feature extraction module, considering the characteristics of unstructured roads and the key points of lightweight model structure design. Based on the TMB module, the overall structure and implementation process of the TMBNet network with an input size of 256×256×3 are presented.

(1) The TMB module was compared with the SS-nbt, Non-bottleneck-1D, and other modules. The effectiveness of the TMB module design was demonstrated through experimental results and a visual display of the test set.
(2) An experiment comparing multiple convolution kernel categories shows that in the TMB module design, using multiple convolution kernels can more fully extract object feature information under different sizes and improve the segmentation accuracy of the network.
(3) A comparison of TMBNet with the accuracy improvement network and existing lightweight semantic segmentation network models further proves the balanced advantage of TMBNet in terms of both network accuracy and forward reasoning speed.

Acknowledgements
The authors express their gratitude to the reviewer for the helpful comments, which improved the manuscript.

Authors' Contributions
FP was in charge of the entire trial; XF supervised the entire work of this paper; DZ and QA wrote the manuscript; RL and JH assisted in writing the manuscript. All authors read and approved the final manuscript.

Funding
Supported by National Natural Science Foundation of China (Grant Nos. 62261160575, 61991414, 61973036), and Technical Field Foundation of the National Defense Science and Technology 173 Program of China (Grant Nos. 20220601053, 20220601030)

Data availability
Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

Declarations

Competing Interests
The authors declare no competing financial interests.

Received: 8 March 2024 Revised: 21 March 2024 Accepted: 27 March 2024
Published online: 17 May 2024

References

[1] Z Chen, L Deng, Y Luo, et al. Road extraction in remote sensing data: A survey. *International Journal of Applied Earth Observation and Geoinformation*, 2022, 112: 102833.
[2] Yuheng Song, Hao Yan. Image segmentation algorithms overview. *arXiv preprint*, arXiv:1707.02051, *2017*.
[3] Tian Zhang, Yong Tian, Zi Wang, et al. Adaptive threshold image segmentation based on definition evaluation. *Journal of Northeastern University (Natural Science)*, 2020, 41(9):1231–1238.
[4] Armin Gruen, Haihong Li. Road extraction from aerial and satellite images by dynamic programming. *ISPRS Journal of Photogrammetry and Remote Sensing*, 1995, 50(4): 11–20.
[5] G Koutaki, K Uchimura. Automatic road extraction based on cross detection in suburb. *Computational Imaging II. International Society for Optics and Photonics*, 2004, 5299: 337–344.
[6] Hong-chun Tan, Li Cai, Ying-bao Geng. An object-based conditional random fields for road extraction. *Remote Sensing Information*, 2016, 31(4): 69–75.
[7] Lifu Chen, Jun Wen, Hongguang Xiao, et al. Road extraction algorithm for high resolution SAR image by fusion of MRF segmentation and mathematical morphology. *Chinese Space Science and Technology*, 2015, 35(2):17–24.
[8] L C Chen, G Papandreou, I Kokkinos, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint, arXiv: 1412.7062, 2014.
[9] Renbao Lian, Weixing Wang, Nadir Mustafa, et al. Road extraction methods in high-resolution remote sensing images: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations And Remote Sensing*, 2020, 13: 5489–5507.
[10] Yanan Wei, Zulin Wang, Mai Xu. Road structure refined CNN for road extraction in aerial image. *IEEE Geoscience and Remote Sensing Letters*, 2017, 14(5): 709–713.
[11] Z Zhang, Q Liu, Y Wang. Road extraction by deep residual U-net. *IEEE Geosci. Remote Sens. Lett.*, 2018, 15(5): 749–753.
[12] P Li, Y Zang, C Wang, et al. Road network extraction via deep learning and line integral convolution. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2016: 1599–1602.

[13] J Xin, X Zhang, Z Zhang, et al. Road extraction of high-resolution remote sensing images derived from DenseUNet. *Remote Sensing*, 2019, 11(21): 2499.
[14] S Y Lo, H M Hang, S W Chan, et al. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. *Proceedings of the ACM Multimedia Asia*, 2019: 1–6.
[15] Y Wang, Q Zhou, J Liu, et al. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019: 1860–1864.
[16] H Zhao, X Qi, X Shen, et al. ICNet for real-time semantic segmentation on high-resolution images. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018: 405–420.
[17] M Denil, B Shakibi, L Dinh, et al. Predicting parameters in deep learning. *arXiv preprint,* arXiv:1306.0543, 2013.
[18] F Chollet. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 1251–1258.
[19] K C Kamal, Z Yin, M Wu, et al. Depthwise separable convolution architectures for plant disease classification. *Computers and Electronics in Agriculture*, 2019, 165: 104948.
[20] A G Howard, M Zhu, B Chen, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint,* arXiv:1704.04861, 2017.
[21] Mark Sandler, Andrew Howard, Menglong Zhu, et al. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *arXiv preprint,* arXiv:1801.04381.
[22] A Howard, M Sandler, G Chu, et al. Searching for mobilenetv3. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 1314–1324.
[23] C Szegedy, S Ioffe, V Vanhoucke, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, https://doi.org/10.1609/aaai.v31i1.11231.
[24] N Ma, X Zhang, H T Zheng, et al. Shufflenet v2: Practical guidelines for efficient CNN architecture design. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018: 122–138.
[25] A Paszke, A Chaurasia, S Kim, et al. ENet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint,* arXiv:1606.02147, 2016.
[26] K He, X Zhang, S Ren, et al. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.
[27] E Romera, J M Alvarez, L M Bergasa, et al. ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2017, 19(1): 263–272.
[28] S Mehta, M Rastegari, A Caspi, et al. ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018: 552–568.
[29] G Lin, A Milan, C Shen, 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.
[30] Z Zhang, Y Pang. CGNet: Cross-guidance network for semantic segmentation. *Science China Information Sciences*, 2020, 63(2): 1–16.
[31] 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, 40(4): 834–848.
[32] H Zhao, J Shi, X Qi, et al. Pyramid scene parsing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2881–2890.
[33] R P K Poudel, U Bonde, S Liwicki, et al. ContextNet: Exploring context and detail for semantic segmentation in real-time. *arXiv preprint,* arXiv:1805.04554, 2018.
[34] R P K Poudel, S Liwicki, R Cipolla. Fast-SCNN: Fast semantic segmentation network. *arXiv preprint,* arXiv:1902.04502, 2019.
[35] C Yu, J Wang, C Peng, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 2018: 325–341.
[36] J Zhuang, J Yang, L Gu, et al. ShelfNet for fast semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, https://doi.org/10.1109/ICCVW.2019.00113.
[37] W Luo, Y Li, R Urtasun, et al. Understanding the effective receptive field in deep convolutional neural networks. *arXiv preprint,* arXiv:1701.04128, 2017.
[38] J Hu, L Shen, G Sun. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 7132–7141.
[39] Q Wang, B Wu, P Zhu, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11534–11542.

Di Zhang is currently a Ph.D. candidate at *School of Automation, Beijing Institute of Technology, China*. His research interests include machine learning and image processing.

Qichao An received the M.S. degree from *Beijing Institute of Technology, China,* in 2021. The main research area is embedded software and hardware development in the field of artificial intelligence.

Xiaoxue Feng received the Ph.D. degree from *Northwestern Polytechnical University, China*, in 2015. She has been a lecturer at *School of Automation, Beijing Institute of Technology, China*, since September 2015. Her research interests include multi-sensor data fusion technology and target detection tracking and recognition.

Ronghua Liu is currently a professional technical expert at the 716th Research Institute of China Shipbuilding Group Co., Ltd. and is a Ph.D. candidate at *Beijing Institute of Technology, China*, where he is mainly engaged in visual measurement, laser measurement, and intelligent algorithm research.

Jun Han received the M.S. degree from *Beijing Institute of Technology, China*, in 2022.

Feng Pan received the Ph.D. degree from *School of Automation, Beijing Institute of Technology, China*, in 2005, and he is currently an associate professor at *School of Automation, Beijing Institute of Technology, China*. His research interests include intelligent information processing and intelligent optimization calculations.