BIR-Net: A Lightweight and Efficient Bilateral Interaction Road Extraction Network

Xianyan Kuang, Fujun Cheng, Cuiqin Wu, Hui Lei, and Zuliang Zhang

***Abstract*—Road segmentation is a crucial aspect in various fields, including intelligent transport systems and urban planning. This article proposes a solution to the problem of inaccurate road region extraction in small devices with limited resources. The proposed solution is a lightweight and efficient bilateral interaction road extraction network, called BIR-Net. First, the detail branch and semantic branch are constructed to form a bilateral feature extraction network for capturing road detail information and semantic information. Then, the shallow interaction module is designed to address the problem of high intraclass variability and interclass similarity in remote sensing images. By exchanging the information of two branches in real time, the edge features of the road are highlighted. The deep interaction fusion module is proposed to fuse information from the two branches using bilateral guided aggregation. Furthermore, to address the characteristics of slender and curved roads in remote sensing images, we have developed the road perception attention module. This module updates the direction weights in real time to track road information, thereby enhancing the network's ability to perceive all road information. The experimental results indicate that BIR-Net has only 3.66M parameters and 6.49G floating point operations. Moreover, the road segmentation accuracies in CHN6-CUG and DeepGlobe datasets are 59.27% and 58.36%, respectively. The proposed method in this article improves road extraction accuracy while maintaining a lightweight structure.**

***Index Terms*—Bilateral interaction network, lightweight and efficient, remote sensing images, road-aware attention module, road extraction.**

I. Introduction

THE rapid advancement of urban modernization and development has led to increased interest in smart city construction, urban planning, and automated driving [1], [2]. Road extraction plays a crucial role in urban planning and decision making [3]. To obtain road information on a large scale, high-resolution remote sensing images have become the primary data source for road area extraction and real-time updates of geospatial databases [4]. Although remote sensing images can be employed to focus on road information on a larger scale, the extraction of roads from these images remains a challenging task due to the irregularity, multiscale nature, high intraclass variability, and high interclass similarity of road information.

Semantic segmentation techniques can be used to distinguish roads from the background in images. The traditional semantic segmentation methods employ techniques, such as morphological and texture analysis for road extraction [5], [6]. These methods require manual operators for feature extraction, followed by template matching and edge detection [7]. The task of road extraction on remote sensing images has been challenging due to the time-consuming manual operations and large errors associated with traditional semantic segmentation techniques.

In light of the accelerated advancement of deep learning, deep convolution-based semantic segmentation neural networks have been put forth and employed for road segmentation and extraction tasks, yielding favorable outcomes. These networks have been successfully employed for road segmentation and extraction tasks. Deep-learning-based semantic segmentation methods typically adhere to an encoder–decoder architectural configuration [8]. The purpose of the encoder is to extract image features layer-by-layer, while the decoder captures the features at different layers and fuses them for pixel classification. FCN [9] is the first to implement semantic segmentation based on the convolutional neural networks, achieving pixel classification through two consecutive convolutional layers. The U-Net family [10], [11] uses a jump-link structure to obtain multiscale features of the road information, making full use of the road information. Similarly, DeepLab V3 [12] expands the sensory field by introducing null convolution in the encoder and uses convolution kernels of different sizes to capture multiscale features. This effectively improves the network's segmentation performance. Additionally, BiseNet V2 [13] and STDC [14] use a multibranch structure to process different information individually and aggregate it, providing better performance in terms of model complexity and segmentation accuracy. Wang et al. [15] extracted features based on context fusion and self-learning sampling to improve segmentation accuracy. Redundant features were effectively suppressed through double feature fusion to reduce the complexity of the network model.

Although these methods based on deep convolutional neural networks demonstrate superior performance in semantic segmentation, they lack comprehensive consideration of local and global information in road features. Furthermore, the extraction

Manuscript received 6 April 2024; revised 8 July 2024; accepted 30 July 2024. Date of publication 9 August 2024; date of current version 19 August 2024. This work was supported by the National Natural Science Foundation of China under Grant 51268017 and Grant 72061016. *(Corresponding author: Cuiqin Wu.)*

Xianyan Kuang, Fujun Cheng, Hui Lei, and Zuliang Zhang are with the School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou 341000, China, and also with the Jiangxi Province Key Laboratory of Multidimensional Intelligent Perception and Control, Ganzhou 341000, China (e-mail: kuangxianyan@jxust.edu.cn; 2500432750@qq.com; 1262290673@qq.com; 1486069198@qq.com).

Cuiqin Wu is with the School of Mechanical and Electrical Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China (e-mail: wucuiqin@jxust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3439267

accuracy of elongated and curved roads in complex remote sensing images remains a challenge.

To enhance the precision of road extraction from remote sensing images, some researchers have utilized large network models as encoders. Zhou et al. [16] expand the convolutional layer in the center of the Linknet [17] architecture and introduced inflated convolution. Zhu et al. [18] incorporate the global context-awareness module into the encoder–decoder to effectively integrate the global context features. Similarly, the authors in [19] and [20] have expanded the model structure by using dilated convolution and ResNet [21] residual connectivity, respectively. These methods enhance the feature extraction capacity of the encoder by incorporating intricate modules, thereby leading to a more pronounced improvement in the accuracy of road extraction from remote sensing images.

However, the adoption of large models and complex modules also presents certain challenges, including augmented model size, elevated number of parameters, more intricate structure, and so forth. Consequently, these models necessitate enhanced computational capacity and arithmetic capability, as well as more sophisticated equipment.

The above network models ignore the significance of shallow information interaction in the context of information fusion, thereby limiting the potential of road information. Furthermore, the interconnection between local and global road information is overlooked in the process of road information enhancement. In order to ensure the extraction accuracy of the network model while simultaneously reducing its weight in order to improve the efficiency of the model, this article proposes the BIR-Net, a lightweight and efficient bilateral interaction road extraction network (BIR-Net). The contributions of this article are as follows.

1) The article proposes a BIR-Net for extracting complex roads in remote sensing images. BIR-Net consists of a detail branch and a semantic branch, which capture road detail information and semantic information, respectively. The two branches are associated and fused using a shallow interaction module (SIM) and a deep interaction fusion module (DIFM). In addition, the proposed road perception attention module (RPAM) enables real-time perception of the location information of narrow and curved roads. Furthermore, the introduction of auxiliary segmentation heads at different feature layers of the semantic branches enhances the multiscale segmentation capability of remotely sensed roads.
2) To address the issue of high intraclass variability and interclass similarity in remote sensing road images, we constructed a gradual-layer interaction module called SIM. The shallow bilateral network learns from each other using maximum pooling and bilinear interpolation methods for upsampling operations, enhancing the network's ability to perceive the intra- and interclasses of the road classes in remote sensing images with respect to detail and semantic information.
3) A proposed DIFM employs a bilateral bootstrap aggregation approach to achieve complementarity between the detail branch and semantic branch information. This enhances the network's overall dependence on road feature extraction, thereby improving the accuracy of road extraction from remote sensing images.
4) An RPAM has been developed to address the challenge of identifying roads with slender curves in remote sensing images. The module constructs horizontal, vertical, and global road networks using average pooling to establish long-term dependencies and capture local road location information. The dependency relationship of the road is captured globally, while local information is sensed in real time. This approach effectively addresses the problem of extracting slender and curved roads in remote sensing images.

II. Related Work

A. Road Network Extraction

Semantic segmentation methods represent a prominent approach for the extraction of road information from remote sensing images. Li et al. [22] propose a hybrid convolutional network that fuses multiple subnetworks to extract road features with different granularities by fusing a full convolutional network, a modified U-net, and a VGG subnetwork. Filho et al. [23] propose an early fusion network with RGB and surface model images to improve road surface extraction by providing complementary geometric data. Tan et al. [24] improve the segmentation accuracy of multiscale roads by encoding and fusing different levels of convolutional layers. Lin et al. [25] propose a dual-task-driven combining road shape patterns constructed as a deep neural network to extract road shape information by residual convolutional coding and bar convolutional decoding. Zhang et al. [26] extract multiscale information by designing a multiscale-assisted predictor and a hybrid loss function. Wu et al. [27] construct a dense residual network to reduce the loss of spatial information and enhance the contextual perception to extract a more complete road region. These methods achieve good performance in road segmentation accuracy in remote sensing images.

However, these road extraction networks are overly concerned with segmentation performance, resulting in a dramatic increase in the number of network parameters, which is severely limited by resources. Therefore, Liu et al. [28] propose a lightweight semantic segmentation model based on the U-Net framework, which reduces the network parameters and improves the computational speed by introducing lightweight modules to replace the convolutional layer in U-Net.

Similarly, the authors in [29] and [30] reduce the complexity of the model by introducing MobileNet V2 and MobileViT modules as the backbone of the network. Liu et al. [31] propose a lightweight road detection network LRDNet based on multiscale convolutional networks and coupled decoding terminals, which increases the parallelism by coupling the decoding processing of threaded tasks and reduces the computational cost. Furthermore, the authors in [13], [32], and [33] use multibranch structure to aggregate information from different layers, which has good real-time performance and accuracy. Zou et al. [34] propose

an all-scale feature fusion network (AF-Net) to extract roads from remote sensing images, which uses an encoder–decoder architecture, and the encoder and decoder are connected by the all-scale feature fusion module. Due to the features with richer semantic information and more precise spatial information, the AF-Net outperforms other methods on two benchmark datasets.

However, these networks only aggregate information from different branches but ignore the importance of shallow information interaction. In contrast, this article adopts the method of shallow interaction and deep interaction fusion to construct a BIR-Net so as to make fuller use of the feature information extracted from detail branch and semantic branch.

B. Enhancement of Road Information Technology

Due to the distinctive characteristics of remote sensing road images, including complex backgrounds and elongated and curved roads, the road feature information enhancement technology of remote sensing images has become a crucial step in improving the accuracy of road segmentation. To extract complete road information, Dai et al. [35] propose a road-enhanced deformable attention network (RADANet) to learn the remote dependencies of road pixels. And a road augmentation module (RAM) is designed to capture the semantic shape information of roads by four-stripe convolution. Lin et al. [25] use a multiscale, multidirectional strip convolution module as a decoder to extract road information using convolution kernels in 0°, 45°, 90°, and 135° directions. This approach is constrained by the availability of local information on road and ignores global information. Dong and Chen [36] design a multidimensional attention module (BMDA) to construct a global attention module using multidimensional information to enhance the global information of the road during feature extraction. The authors in [18] and [37] enhance the global dependence of the network on the road by integrating contextual features to capture global information.

In addition, Luo et al. [38] introduce a hybrid receptive field module in the encoder to enhance the target road features by adaptively adjusting the receptive field sizes of roads at different scales. Similarly, the authors in [20], [39], and [40] enhance the receptive field of remotely sensed roads by introducing the dilated convolution and inflated convolution attention modules. The edge information of the road is important to distinguish the road from the background; Ge et al. [41] propose a new feature viewing transmission network. It mitigates the phenomenon of broken road segments and missing connections in road extraction by improving the contour learning ability. Wang et al. [42] improve the feature extraction ability of the network when the road is occluded by designing the dilated attention across stages. Qiu et al. [43] propose the feature refinement module to refine the road texture and detail information.

The aforementioned methods merely refine road information in a unilateral manner, ignoring the connection between local and global information. In contrast, the RPAM proposed in this article is capable of capturing both local road information and global road dependencies in real time.

III. Methodology

In this section, the framework of the bilateral interaction road extraction method is described in detail. There are two branches of the framework, one is the detail branch for extracting the detail information of the road, and the other is the semantic branch for extracting the semantic information of the road. The shallow detail branch and the semantic branch information are interacted through SIM to guarantee the completeness of the road information extraction, and DIFM is used to aggregate the feature information of the two branches. Furthermore, in consideration of the elongated and curved characteristics of roads, the RPAM proposed in this article enhances the continuity of the road extraction region by detecting the road location in real time and capturing the global dependency of the road.

A. Network Framework

The framework of BIR-Net network is shown in Fig. 1. The network framework mainly consists of three modular parts: detail branch, semantic branch, and interaction fusion. The first module is the detail branch, in which multiple 3×3 convolutions are stacked, and the downsampling operation is performed using a 3×3 convolution with a step size of 2. This branch is distinguished by its broad channels and shallow layers, which are employed to capture low-level details and generate high-resolution feature representations.

The second module is the semantic branch, which can effectively reduce the computational cost by narrowing down the features and fusing them through two different downsampling operations, thereby reducing the number of calculations required. In the subsequent downsampling, in order to ensure that the computational load is reduced while the receptive field is increased, two 3×3 depth separable convolutions are used instead of the 5×5 depth convolution, and a 1×1 convolution is used as a projection layer to project the output of the depth convolution into the low channel capacity space. The feature representation of this branch is distinguished by narrow channels and deep layers to capture high-level semantic information. Meanwhile, an auxiliary segmentation header [13] is introduced in different layers of semantic branch to extract multiscale road information. The segmentation header consists of a convolution with a 3×3 convolution kernel, batch normalization, ReLU nonlinear activation function, and a convolution operation with a 1×1 convolution kernel, with the output dimensions set to the number of semantic categories to be segmented. The RPAM is added at the last layer of this branch to improve the road perception capability of the network.

The third module is the interaction fusion module, which consists of an SIM and a DIFM, respectively. The SIM module employs maximum pooling and bilinear interpolation to enable the communication learning of features in detail and semantic branches, thereby enhancing the network's ability to perceive remote sensing images within and between classes. DIFM achieves the function of complementing detail branch and semantic branch information by means of bilateral bootstrap aggregation.

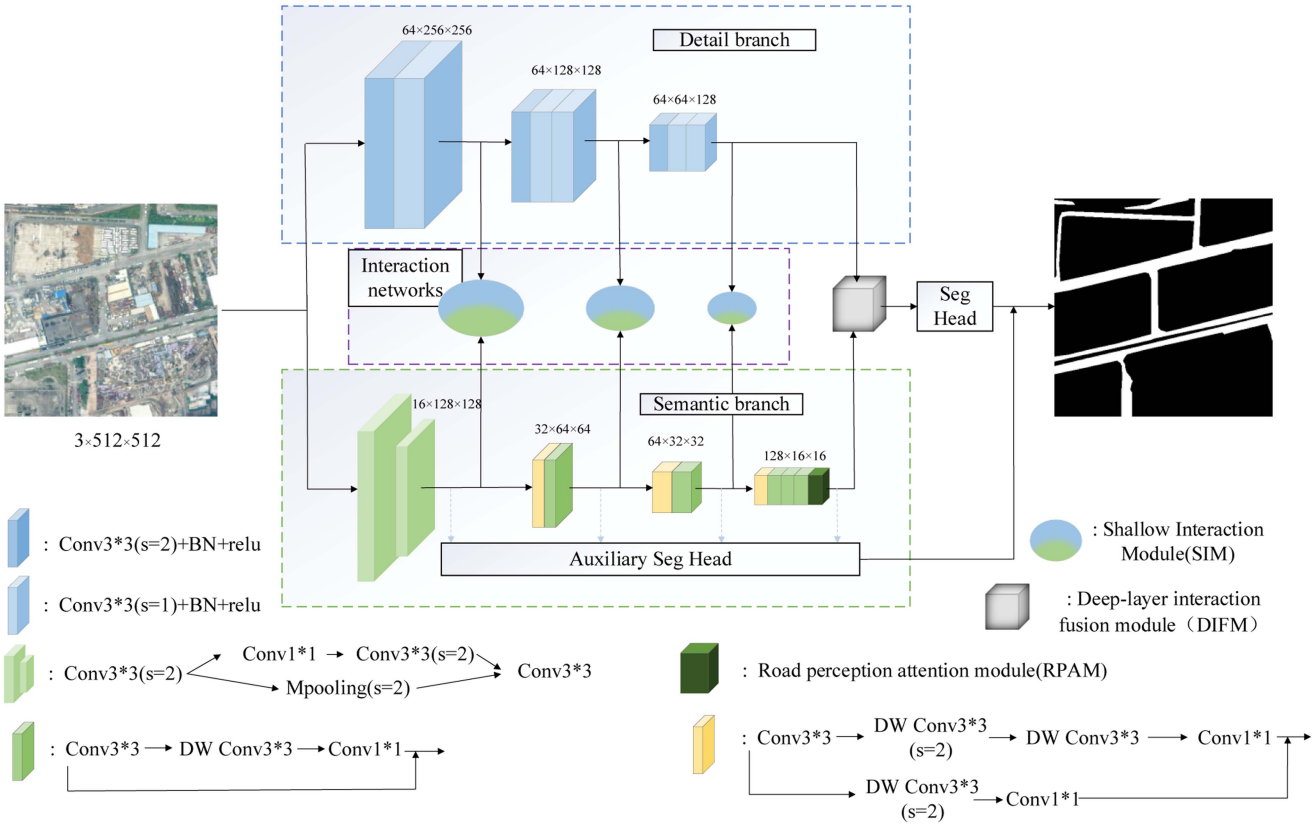

Fig. 1. Framework of the proposed BIR-Net network. The segmented backbone network has a detail branch represented by the blue dotted box, a semantic branch represented by the green dotted box, a shallow interaction network represented by the purple dotted box, and a deep interaction fusion network represented by the gray module. The network first extracts detailed and semantic information from the picture using a two-branch backbone network. The shallow information in the two branches is learned interactively during the extraction process. The feature information of the two branches is then fused in the last layer. Finally, a segmentation header distinguishes the road from the background. The convolutional module's operation details are described below the network.

The BIR-Net employs a two-path network architecture, comprising a spatial path (detail branch) and a contextual path (semantic branch). The architecture permits the network to concentrate on both spatial details and contextual information, circumventing the intricacies of traditional approaches that necessitate the processing of both. This architecture effectively reduces the number of network parameters while maintaining the performance of the network. Furthermore, the SIM is designed to be completed between neighboring layers when interacting with spatial and semantic information. This approach avoids cross-layer information fusion in traditional methods, effectively reducing the computational burden on network models. Consequently, BIR-Net achieves a lightweight design while maintaining high performance.

B. Shallow Interaction Module

Remote sensing images are distinguished by high intraclass variability and high interclass similarity. Specifically, the challenge arises primarily from the significant discrepancies in the characteristics exhibited by entities belonging to the same semantic class. The variability in the style, shape, size, and distribution of entities frequently presents challenges in accurately extracting roadway information. The issue of high interclass similarity can be attributed primarily to the presence of identical objects that overlap between different scene classes or high semantic scene categories.

Currently, many segmentation networks [23], [44], [45] are prone to the phenomenon of "disconnection" in the feature extraction stage; in other words, it is difficult to ensure that the details and semantic information of each layer are compatible. This leads to the inability to achieve good results in the remote sensing road extraction task. Therefore, in this article, an SIM is designed, as shown in Fig. 2. The SIM module facilitates the communication of both detail and semantic information while extracting remote sensing road features. This enables layer-by-layer information to learn from each other and to monitor the changes in features within and between classes of remote sensing images.

The issue of significant intraclass variability pertains to the fact that roads may exhibit considerable dissimilarities in their spectral characteristics. These dissimilarities may originate from disparate materials, age, maintenance status, and other variables. In such instances, algorithms may encounter difficulties in discerning road areas exhibiting disparate spectral characteristics as belonging to the same class when relying exclusively on spectral data. In order to address this issue, it is essential to consider the role of spatial information. Spatial information encompasses

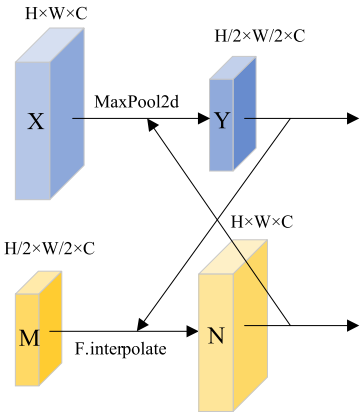

Fig. 2. SIM. The feature map for the detail branch is represented in blue, while the feature map for the semantic branch is represented in yellow. The network facilitates interactive learning by merging the feature information from both branches.

not only the location of individual pixels but also the spatial relationships between them, as well as the shape, direction, and width of the road. Accordingly, in order to utilize the spatial information in the detail branch, SIM operates as follows.

First, a maximum pooling layer (MaxPool 2-D) with a convolution kernel size of 3×3 and a step size of 2 is selected for the downsampling operation in the detail bifurcation layer, which can be computed as follows:

$$Y[i,j] = \max(x[p,q]) \tag{1}$$

where $Y[i, j]$ is the value of the pixel in row i and column j of the output feature map, x is the input feature map, p and q are the row index and column index of the input feature map, and max is the operation of taking the maximum value.

The method reduces the feature map x of size $H\times W$ to the feature map y of size $H/2 \times W/2$ based on the number of feature channels that remain unchanged, with Y retaining the texture information in the detail branch.

Then, the feature map Y and the feature map M in the semantic branch are add fused so that the semantic branch successfully receives the detail information that is more sensitive to the texture feature information.

The term "high interclass similarity" refers to the fact that roads are in close proximity to other features (e.g., rivers, bare soil, etc.) in terms of their spectral properties. This makes it challenging to accurately distinguish between them based solely on pixel-level spectral information. In this instance, it is of paramount importance to consider the contextual information. Contextual information encompasses information about the local area surrounding the pixel, the broader regional context, and the topology of the road network. This information furnishes the algorithm with spatial relationships and dependencies between pixels, thereby facilitating more accurate recognition of roads. Consequently, in order to utilize the contextual information in the semantic branch, the processes of SIM are as follows.

For the input size of $H/2\times W/2$ feature map M, the semantic branching layer selects the bilinear interpolation method for upsampling to obtain the feature map N of size $H\times W$, and the computational steps are shown as follows:

$$f(R_1) = \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21})$$

$$f(R_2) = \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}) \tag{2}$$

$$f(P) = \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2) \tag{3}$$

where $Q_{11} = (x_1, y_1)$, $Q_{12} = (x_1, y_2)$, $Q_{21} = (x_2, y_1)$, and $Q_{22} = (x_2, y_2)$ are the four closest points to point P, which is the target pixel point $f(x, y)$.

The linear interpolation in the x-direction is calculated by (2) to obtain $f(R_1)$ and $f(R_2)$, which is inserted into (3) to calculate the linear interpolation $f(P)$ in the y-direction, and finally, the target pixel value P is obtained by simplification, as shown in the following equation:

$$\begin{aligned} f(P) = f(x,y) = \frac{1}{(x_2-x_1)(y_2-y_1)} [&f(Q_{11})(x_2-x)(y_2-y) \\ &+ f(Q_{21})(x-x_1)(y_2-y) + f(Q_{12})(x_2-x)(y-y_1) \\ &+ f(Q_{22})(x-x_1)(y-y_1)]. \end{aligned} \tag{4}$$

This method allows for the retention of high semantic information, and the incorporation of feature information N into the detail branch via the ADD method enables the detail branch and the semantic branch to communicate successfully.

The utilization of SIM can effectively address the issue of high intraclass variability and high interclass similarity in remote sensing images by integrating contextual and spatial information.

C. Deep-Layer Interaction Fusion Module

Traditional fusion methods [46], [47] combine two types of feature responses by summation and concatenation in an elementary way. The low-layer detail branch and the high-layer semantic branch exhibit disparate levels of feature representation, so the simple combination of these two types of information is ineffective in harnessing their full potential. Yu et al. [13] have inherently encoded multiscale information by capturing feature representations at different scales and allowing simple complementation of the two branches of information.

In this article, a DIFM is constructed, as shown in Fig. 3. Bilateral bootstrap aggregation is employed in the DIFM module to achieve a complementary fusion of deep information from the detail branch and the semantic branch. In brief, the detail branch features are downsampled and interact with the semantic branch features, while the semantic branch features are upsampled and interact with the detail branch features. This deep-layer complementarity allows information from both branches to be fully utilized through interaction fusion.

The precise location of the road within the deep interaction fusion is of lesser importance than the relative location of the road in relation to the background. Therefore, first, in the detail branch, downsampling is performed using two times of average pooling in order to obtain feature information that is more

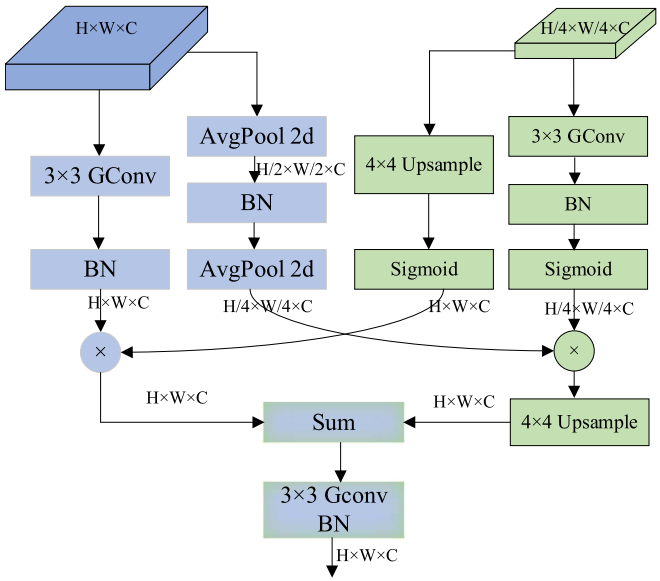

Fig. 3. DIFM. The network's blue section represents the steps taken during the fusion of the detail layer, while the green section represents the steps taken during the fusion of the semantic layer. The fusion process involves two layers of interaction.

sensitive to the background information, which is beneficial for the subsequent classification task.

Then, to enhance the precision of the road segmentation algorithm while maintaining a lightweight model, the ghost convolution (GConv) [48] is introduced into the DIFM module to replace the normal convolution. This convolution replaces the regular convolution approach by combining a small number of convolution kernels with a cheaper linear variation operation to achieve higher model performance with lower computational cost, and the ghost convolution is calculated as shown in the following equation:

$$\begin{aligned} y_1 &= \text{Conv}(x) \\ y_2 &= \Phi_{i,j}\left(y_1\right) \quad \forall i = 1, \ldots, m, \quad j = 1, \ldots, s \\ Y &= \text{concat}(y_1, y_2) \end{aligned} \tag{5}$$

where y_1 is the feature map of the current evidence using regular convolution, representing linear transformations in different directions. The feature maps y_1 and y_2 are fused by concatenation to obtain the fused feature map Y. Φ denotes a convolution operation on a feature-by-feature layer.

Finally, the information in the two branches is summed and the features are further fused by a convolutional kernel size of 3×3 ghost convolution, resulting in a feature map that is more conducive to the road segmentation task.

D. Road Perception Attention Module

The target features in traditional images are predominantly distributed in a block-like manner. However, the bar-shaped roads in remote sensing images exhibit a distinctive shape. Therefore, the traditional square convolution kernel gets more irrelevant information when extracting road features. Consequently, as long as the location information of the strip road is perceived, the road shape in the specified direction can be taken into account, thereby enhancing the accuracy of the road feature extraction.

Road features can be captured more efficiently by using attention mechanisms [29], [31]. The more classical channel attention mechanisms include squeeze-and-excitation (SE) attention [49] and convolutional block attention module (CBAM) [50]. Coordinate attention (CA) [51] decomposes channel attention into a 1-D feature encoding process in two spatial directions, which can capture long-range dependencies in one spatial direction while preserving precise position information in the other. However, it is difficult for CA to capture the global information of roads with elongated curved features in remote sensing images.

In this article, a new RPAM is constructed, of which the structure is shown in Fig. 4. The RPAM module consists of three branches: horizontal branch, vertical branch, and global branch. The horizontal branch and the vertical branch are used to capture the location information of the road, and the global branch captures the global information.

First, for the input P, a pooled convolution kernel with dimensions $(H, 1)$ $(1, W)$ is used to encode each channel along the horizontal and vertical coordinates, respectively, to obtain the features Z_h and Z_w, which are aggregated along the two spatial directions. To ensure that the information of the two spatial directions interacts with each other, the information of the two is fused using a concat operation, defined as follows:

$$Z_h = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i)$$

$$Z_w = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{6}$$

$$F = \text{concat}(Z_h, Z_w) \tag{7}$$

where H and W are the height and width, respectively, and F is the fused feature map.

Then, the tensor is decomposed along the spatial dimensions by the split function, thereby yielding the feature maps X and Y, respectively. Meanwhile, considering the characteristics of slender and curved roads in remote sensing images, the 3×3 convolution and sigmoid activation functions are used to capture finer road features. The generation of attentional weights for height and width is achieved through the splitting process and the application of a sigmoid activation function. The weights are employed to reweight the height and width pooled feature maps, respectively, with the objective of emphasizing spatial road regions and suppressing background regions. By multiplying the original feature maps X and Y with the reweighted height and width attentional feature maps, the model is made more spatially focused on road areas. The operation is defined as follows:

$$\begin{aligned} g^h &= \sigma\{s[\text{conv}_{3\times3}(F_h)]\} \times X \\ g^w &= \sigma\{s[\text{conv}_{3\times3}(F_w)]\} \times Y \end{aligned} \tag{8}$$

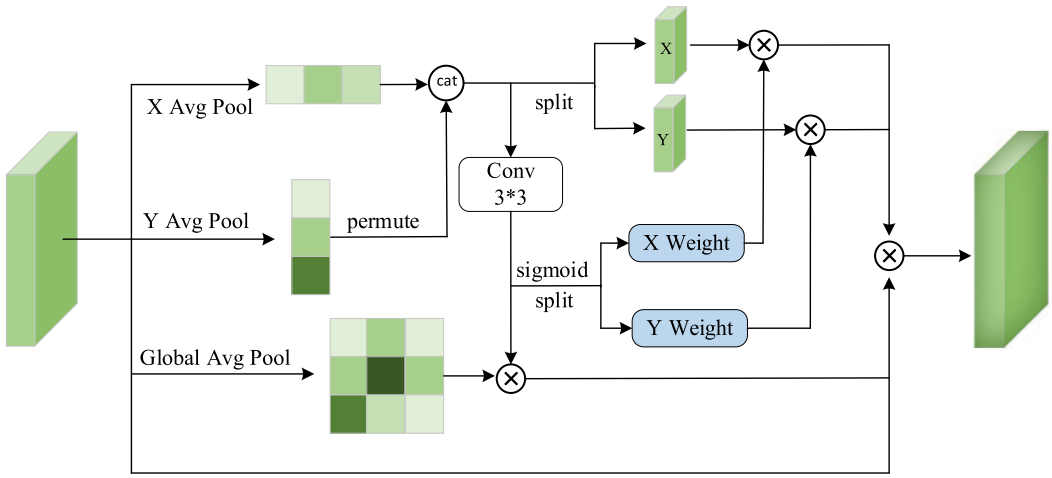

Fig. 4. Structure of RPAM. The network comprises three branches: the horizontal branch (horizontal grid), the vertical branch (vertical grid), and the global branch (square grid). The blue sections indicate the weights of the road information in the horizontal and vertical directions.

TABLE I
EXPERIMENTAL ENVIRONMENT

Items	Details
System	Windows11
CPU	R7-7735H，(3.2 GHz)
RAM	16 GB
Graphics Card	GeForce RTX4050
DL Framework	Pytorch1.10.2.
Optimizer selects	0.01
Decay Coefficient	0.01
Epoch	200

where s is the sigmoid activation function and σ denotes the split function.

Furthermore, the input feature maps are globally average pooled by concatenating a global average pooling branch, which generates the average value for each channel and, subsequently, generates the channel attention weights. Then, the resulting weights are normalized to a value between 0 and 1 by means of a sigmoid activation function, thereby indicating the relative importance of each channel. By multiplying the original feature map with the channel attention weights, the model can focus on or suppress different channels, which helps to emphasis the channels that contain road information.

Finally, the feature information from all branches is fused. By multiplying the height, width, and channel attention weights with the original feature map, the RPAM is capable of effectively capturing road features in remote sensing images by considering both channel and spatially significant information.

IV. EXPERIMENTS

A. Datasets

CHN6-CUG dataset [52]: The CHN6-CUG dataset is of remote sensing images of urban roads in China, with images from six cities, namely Beijing, Shanghai, Wuhan, Shenzhen, Hong Kong, and Macau. The spatial resolution of the images is 50 cm/pixel. CHN6-CUG contains 4511 labeled images of 512×512 pixels, of which 3608 are used for model training and 903 for testing.

DeepGlobe dataset [53]: DeepGlobe dataset contains 6226 satellite remote sensing images of 1024 × 1024 pixels and labels, each with a spatial resolution of 50 cm/pixel. The images contain urban, suburban, and rural roads from Thailand, India, and Indonesia, of which 4980 are used for model training and 1246 for testing.

B. Experimental Setting and Evaluation Indicators

The experimental environment is shown in Table I.

The SGD is selected as the model training optimizer with an initial learning rate of 0.01 and a learning rate decay coefficient of 0.01. The training period is set to 200 rounds with four images per batch.

In order to comprehensively evaluate the lightweight and segmentation performance of the model, Param (number of parameters), floating point operations (FLOPs), and RIoU are used as the evaluation metrics. Among them, Param and FLOPs are used to evaluate the size and complexity of the model, and RIoU is the intersection ratio of road categories. They are explained in detail as follows.

1) Param is the number of parameters required for model training, which can measure the computational complexity of the model.
2) FLOPs are the number of floating point operations, which can measure the computational time complexity of the model.
3) RIoU can measure the segmentation effect of the model on the road category, which is the intersection ratio of the true and predicted values of the road category, and is calculated as follows:

$$\mathrm{RIoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

TABLE II
BIR-NET ABLATION EXPERIMENTS

Model	Param (M)	FLOPs (G)	Road IoU(%)	
			CHN6-CUG	DeepGlobe
DSNet+Sum	**2.98**	**5.95**	56.23	56.19
DSNet +DIFM	3.42	6.26	57.37	56.47
DSNet +DIFM+SIM	3.57	6.48	58.83	58.04
DSNet +DIFM+SIM+RPAM (BIR-Net)	3.66	6.49	**59.27**	**58.36**

The results in bold denote the best performance, and underline denote the second-best performance among different methods.

TABLE III
EXPERIMENTS VALIDATING TECHNOLOGY EFFECTIVENESS

Feature extraction	Feature fusion Method	Feature Enhancement	Param(M)	FLOPs(G)	RIoU(%)
DSNet	Sum	-	**2.98**	5.95	56.23
	DIFM (without GConv)		3.41	**6.57**	57.03
	DIFM		3.42	6.26	57.37
	Sum	SE	3.15	5.95	57.68
		CA	3.01	5.95	57.28
		CBAM	2.99	5.96	56.89
		BAM	2.99	5.95	56.95
		RPAM	3.07	5.96	57.89
DSNet+SIM	Sum	-	3.13	6.17	58.08
	DIFM	**RPAM**	3.66	6.49	**59.27**

The results in bold denote the best performance, and underline denote the second-best performance among different methods.

where TP stands for true example, FP stands for false positive example, and FN stands for false negative example.

C. Experimental Results

1) Ablation Experiments: Ablation experiments are conducted to evaluate the impact of each module in the BIR-Net on the network segmentation performance. To facilitate the experiments, we define DSNet (network with detail branch and semantic branch without RPAM) as the basic bilateral feature extraction network. The sum operation of two DSNet branches has been defined. The features of the detail and semantic branches in DSNet are fused by a simple sum operation, which is designated as DSNet+Sum.

The DSNet network is utilized as the base network, with the addition of DIFM, SIM, and RPAM modules to form the DSNet+Sum+DIFM, DSNet+Sum+DIFM+SIM, and DSNet+ Sum+DIFM+SIM+RPAM models. The validation is performed on two datasets, CHN6-CUG and DeepGlobe, and the experimental results are presented in Table II.

The experimental data show that the model parameter count (Param) and the FLOPs for road extraction by DSNet+Sum only are 2.98M and 5.95G, respectively. The road IoU reaches 56.23% and 56.19% on CHN6-CUG and DeepGlobe datasets, respectively. The addition of DIFM module only increases Param and FLOPs by 0.44M and 0.31G, respectively. The RIOU reached 57.37% and 56.47% on CHN6-CUG and DeepGlobe datasets, respectively. This indicates that richer details and semantic information are employed to optimize model performance.

Additionally, the SIM is added to maintain real-time connections between detail and semantic branches. As a result, road segmentation accuracy improves to 58.83% and 58.04% for detail and semantic branches, respectively. Finally, the RPAM is included to detect road location features in real time, resulting in a further improvement of model performance by 0.44% and 0.32%, respectively. The DSNet+ DIFM+SIM+RPAM, which is the proposed BIR-Net, has only Param of 3.66M and FLOPs of 6.49G. The road extraction accuracies are up to 59.27% and 58.36%, respectively.

2) Technical Validity Experiments: To validate the effectiveness of the proposed BIR-Net in this article, we compare the proposed modules and some general techniques on the CHN6-CUG dataset. The experimental results are shown in Table III, in which DSNet is used as the base model for validation. The proposed method in this article is represented by the bolded part, while sum is a simple additive feature fusion.

The use of DIFM instead of simple sum fusion results in an improvement of 0.8% in model accuracy. Additionally, the introduction of ghost convolution (GConv) reduces the model complexity by 0.31G and improves accuracy by 0.34% with minimal increase in the number of parameters. The classical feature enhancement modules SE and CA attention can improve features by constructing channel global dependence and location information, respectively. This results in an effective accuracy improvement of 1.45% and 1.05%, respectively. CBAM and BAM combine spatial attention and channel attention mechanisms through series–parallel connection, resulting in a slight improvement in model accuracy.

However, the aforementioned enhancement modules are constrained in their capacity to accurately capture the long-range dependence of roads due to the specificity of road shape features in remote sensing images. In contrast, the RPAM proposed in this article is capable of dynamically focusing on the road location information in real time based on the road feature weights. This

TABLE IV
COMPARISON OF MAINSTREAM MODELS

Model	Param(M)	FLOPs(G)	Road IoU(%)	
			CHN6-CUG	DeepGlobe
U-Net [10]	28.09	78.19	49.53	49.73
Deeplab V3 (ResNet18) [12]	13.60	42.99	53.76	54.26
DDRNet [57]	20.29	8.94	57.63	57.45
D-LinkNet [16]	29.5	43.56	55.87	57.25
STDC [14]	14.23	23.51	58.23	57.14
PSPNet(MobileNet V2) [55]	2.64	9.86	52.17	54.84
Mobile ViT [56]	**1.21**	**0.42**	42.92	47.73
BiseNet V2 [13]	3.62	6.40	56.42	56.08
RADANet [35]	73.58	212.0	**60.51**	**59.67**
RoadViT [54]	1.25	1.18	57.0	52.3
PIDNet [33]	7.83	6.52	58.46	58.17
PP-LiteSeg [58]	5.12	3.21	57.56	57.14
BIR-Net(ours)	3.66	6.49	59.27	58.36

The results in bold denote the best performance, and underline denote the second-best performance among different methods.

leads to an effective improvement in the model's accuracy, with an increase of 1.66%.

Furthermore, to substantiate the functionality of the SIM, an investigation is conducted to ascertain the efficacy of the SIM module when introduced in isolation. The results demonstrate a significant improvement in model accuracy, with an increase of 1.85%. The ability of SIM to facilitate the interaction between detail branch and semantic branch information, as well as to attend to both intra- and interclass information in remote sensing images, has been demonstrated to increase the accuracy of the model.

3) Comparison of Mainstream Models: To verify the advancement of BIR-Net, we compare it with mainstream road segmentation models. The large models compared include U-Net [10], DeepLab V3 (ResNet18) [12], DDRNet [57], D-LinkNet [16], STDC [14], and RADANet [35]. The small models compared include PSPNet (MobileNet V2) [55], MobileViT [56], BiseNet V2 [13], RoadViT [54], PIDNet [33], and PP-LiteSeg [58]. Experiments are conducted using the CHN6-CUG and DeepGlobe datasets, and the results are presented in Table IV.

To ensure the experiment's validity, the loss function and accuracy curves of each model converged within 200 epochs of training, as shown in Fig. 5. The experimental curves indicate that all models have achieved convergence in both their loss function and accuracy.

Table IV presents that our BIR-Net model achieves a road segmentation accuracy of 59.27% on the CHN6-CUG dataset. Our BIR-Net model has a parameter count of only 3.66M and a complexity of 6.49G while still being sufficiently lightweight. Compared with larger models, such as U-Net, Deeplab V3 (ResNet18), and D-LinkNet, the BIR-Net is more advantageous in terms of both accuracy and lightness.

Meanwhile, DDRNet, PIDNet, and STDC improve the segmentation accuracy by constructing a multibranch aggregation structure to eliminate redundant channels and parameters in the network structure. However, the performance of these models in road extraction accuracy remains inferior to that of BIR-Net.

RADANet achieves the highest road extraction accuracy by using a large encoder and decoder as the base network and capturing the road shape information through the road enhancement module. However, pursuing segmentation accuracy excessively by using a deep encoding–decoding structure leads to an overly complex network with a dramatic increase in the number of parameters and computational complexity. The Param and FLOPs of RADANet are as high as 73.58M and 212.0G, respectively, which are the highest among all models. In comparison, BIR-Net has only 4.98% of the Param of RADANet and only 3.06% of FLOPs of RADANet, making it more suitable for deployment on smaller devices.

In comparison with the lightweight models Mobile ViT and RoadViT, BIR-Net exhibits a slightly higher number of parameters and complexity; however, it has been demonstrated that BIR-Net improves model accuracy by 16.35% and 2.77%, respectively. PSPNet(MobileNet V2) uses a lightweight network as the backbone to reduce the number of parameters, but the accuracy is also decreased. BIR-Net improves accuracy by 7.1% with lower model parameter number and complexity than the former by 1.02M and 3.77G, respectively. PP-LiteSeg employs a flexible and lightweight decoder to streamline the encoding–decoding network structure, effectively achieving a balance between segmentation accuracy and lightweight. However, the segmentation accuracy of this approach is inferior to that of BIR-Net.

Furthermore, BIR-Net demonstrates a higher segmentation accuracy compared with its predecessor, BiseNet V2, by 2.88% in cases where the number and complexity of parameters are comparable. Based on these analyses, it is evident that BIR-Net is a remote sensing image road extraction model that balances model complexity and accuracy.

To validate the effectiveness of BIR-Net, this article conducts additional comparison experiments on the DeepGlobe dataset. Table IV presents a comparison between the lightweight models PSPNet (MobileNet V2) and BiseNet V2, which have similar numbers of parameters and complexity. The segmentation

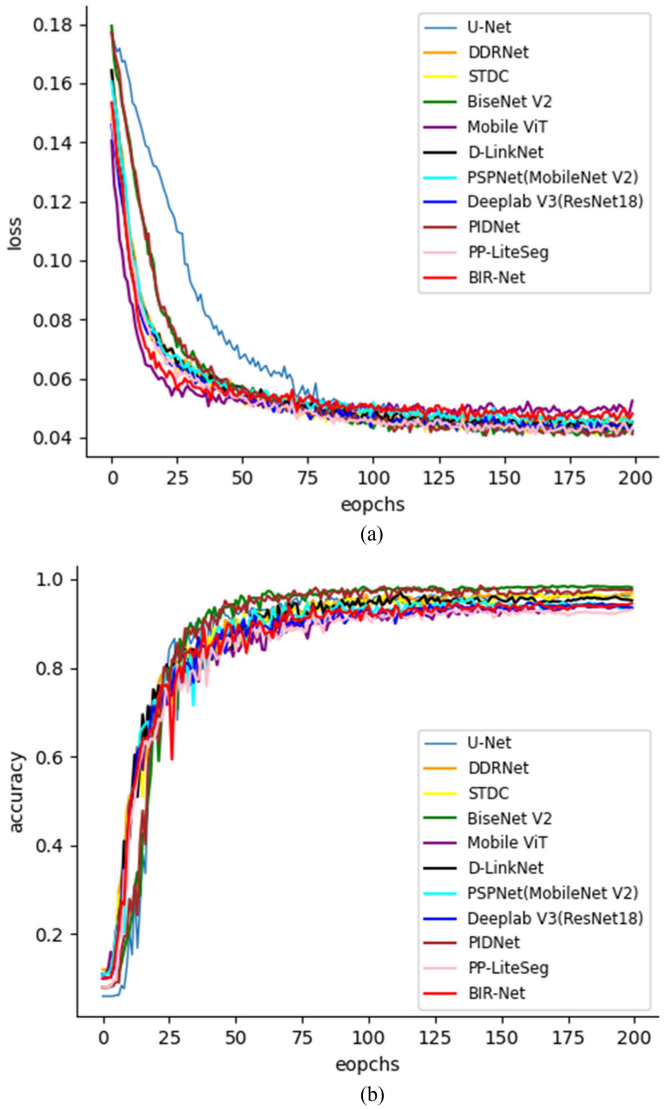

Fig. 5. Loss function curves and accuracy curves for each model. The horizontal axis represents the number of training epochs, while the vertical axis represents the values of the loss function and accuracy. As training time increases, the loss function and accuracy curves gradually converge. (a) Loss function curves. (b) Accuracy curves.

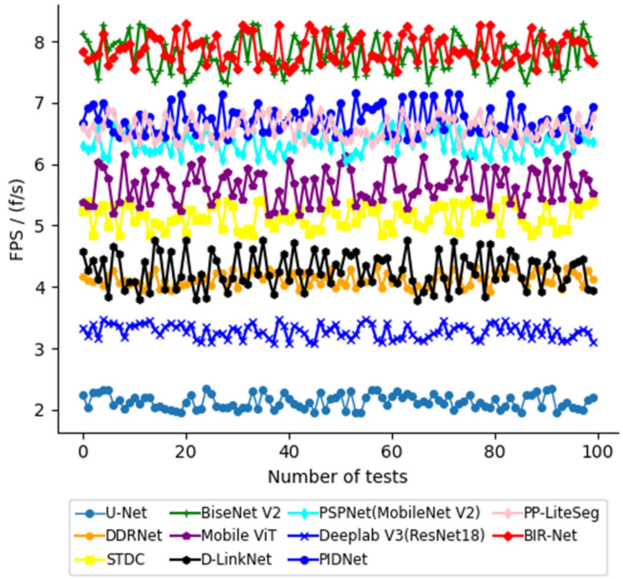

Fig. 6. Comparison of processing time for each model. The figure represents the testing process for each model using different colors. The horizontal coordinate indicates the number of tests, and the vertical coordinate indicates the speed of testing.

accuracies of BIR-Net are higher than those of the former by 3.52% and 2.28%, respectively. However, Mobile ViT and RoadViT have lower numbers of parameters and complexity, resulting in model accuracies of only 47.73% and 52.3%, respectively, which do not take into account the segmentation accuracies. However, Mobile ViT and RoadViT have lower numbers of parameters and complexity, resulting in model accuracies of only 47.73% and 52.3%, respectively, which do not take into account the segmentation accuracies. However, Mobile ViT and RoadViT have lower numbers of parameters and complexity, resulting in model accuracies of only 47.73% and 52.3%, respectively, which do not take into account the segmentation accuracies. PIDNet exhibits a similar degree of model accuracy to BIR-Net, yet it comprises more than twice the number of parameters and is less effective than BIR-Net in terms of model lightweighting. PP-LiteSeg demonstrates an acceptable level of model complexity, yet its segmentation accuracy is notably inferior.

Compared with the larger models, such as U-Net, Deeplab V3 (ResNet18), DDRNet, D-LinkNet, and STDC, BIR-Net offers a more advantageous balance between lightness and accuracy. RADANet achieves the highest segmentation accuracy, but its complexity, with a parameter count of 73.58M, makes it unsuitable for implementation on small devices for road extraction tasks. In conclusion, the proposed BIR-Net model is capable of effectively combining segmentation accuracy with a lightweight design. It has the advantages of being a small model with high accuracy for the mainstream remote sensing image road extraction task.

V. Discussion

A. Evaluation of the Processing Speed

In order to assess the efficiency of BIR-Net, the comparison is made with mainstream models in terms of processing time. The test is conducted on a 512×512 image, and the results are presented in Fig. 6.

Fig. 6 demonstrates that the BIR-Net outperforms the mainstream models in terms of processing speed, with an average FPS of 8 $f \cdot s^{-1}$, and is placed in the leading position. Although MobileViT and PSPNet (MobieNet v2) perform well in terms of model complexity, their average processing speeds are only 5.5 $f \cdot s^{-1}$ and 6.2 $f \cdot s^{-1}$, respectively.

DDRNet and STDC perform moderately well in terms of segmentation accuracy as well as lightness. D-LinkNet and DDRNet have similar processing speeds, with an average FPS of only 4.2 $f \cdot s^{-1}$. Furthermore, U-Net employs a complex

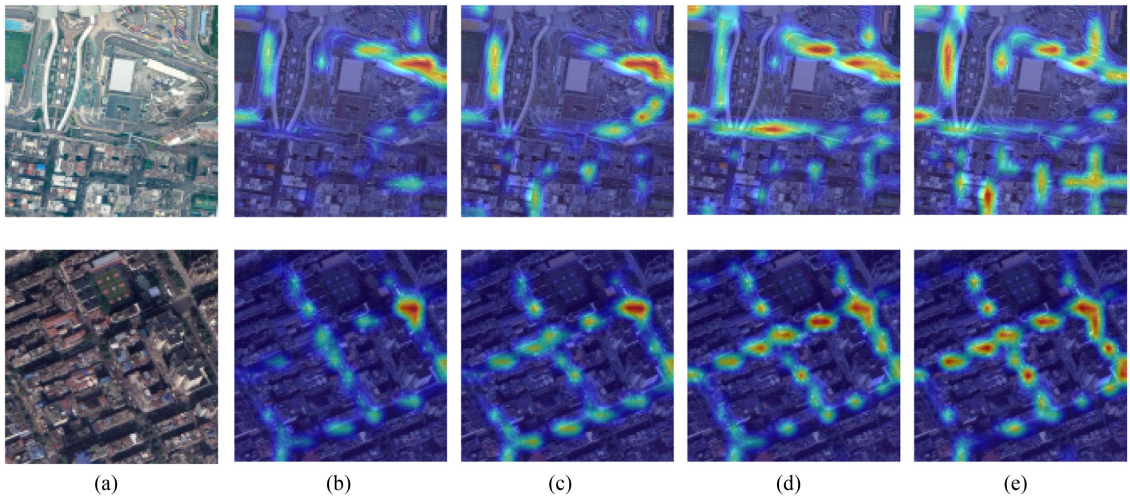

Fig. 7. Comparison between the intermediate result output heat maps. The brightness of the illuminated area indicates the level of attention the network gives to the road features. The proposed BIR-Net pays the most attention to roads, especially with the inclusion of modules. (a) Original image. (b) DSNet+Sum. (c) DSNet+DIFM. (d) DSNet+DIFM+SIM. (e) BIRNet (Ours).

encoding–decoding structure to extract features, while Deeplab V3 (ResNet18) uses the large ResNet18 model for feature extraction. However, their processing speeds are only about 2.3 $f{\cdot}s^{-1}$ and 3.4 $f{\cdot}s^{-1}$, respectively.

PIDNet and PP-LiteSeg reduce the complexity of the model by adopting a multipath network frame and a flexible and lightweight decoder, respectively. The average processing speed reaches 6.8 $f{\cdot}s^{-1}$ and 6.6 $f{\cdot}s^{-1}$, respectively. However, this is still not as good as that of BIR-Net. BiseNetV2 utilizes a concise and effective model structure for feature extraction, which is similar to BIR-Net in terms of processing speed. However, BiseNetV2 performs poorly in terms of road segmentation accuracy.

In summary, the proposed BIR-Net achieves an average FPS of 8 $f{\cdot}s^{-1}$ and has higher processing speed than other mainstream models on images with a size of 512×512. This further validates the lightweight nature of our BIR-Net.

B. Evaluation of Ablation Experiments Effect

To demonstrate the process of feature extraction by BIR-Net, two images were selected at random to generate a feature heat map, as shown in Fig. 7.

Fig. 7(a) shows the input image, Fig. 7(b) shows the feature heat map of the bilateral fusion base network, and Fig. 7(c)–(e) shows the effect diagrams of adding DIFM, SIM, and RPAM in sequence.

Following the replacement of the sum feature fusion with DIFM, the road information is enhanced and a greater number of road areas are illuminated. The incorporation of SIM has resulted in enhanced focus and clarity of the edges of the learned road features. Following the incorporation of RPAM, the global information pertaining to the road is augmented, culminating in the generation of a comprehensive and lucid road feature effect map. This outcome serves to substantiate the efficacy of BIR-Net in the extraction of road features.

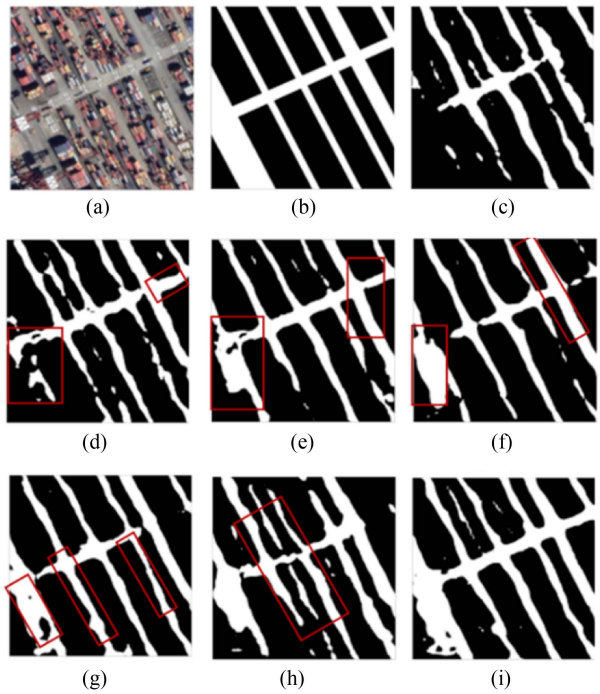

Fig. 8. Impact of segmentation effects from different techniques. The differences are boxed in red, DFM improves the result completeness, RPAM improves the accuracy, and SIM improves the connectivity. The proposed BIR-Net achieves optimal performance. (a) Original image. (b) Real label. (c) DSNet+sum. (d) DSNet+DIFM (without GConv). (e) DSNet+DIFM. (f) DSNet+Sum+RPAM. (g) DSNet+Sum+SIM. (h) BIR-Net (without SIM). (i) BIR-Net.

C. Discussion on Technical Effects

In order to further validate the impact of the proposed model on the performance of road extraction, binarized images are employed to facilitate a comparative analysis of the effects of different techniques. The results are shown in Fig. 8, where Fig. 8(a) shows the original image and Fig. 8(b)–(h) shows

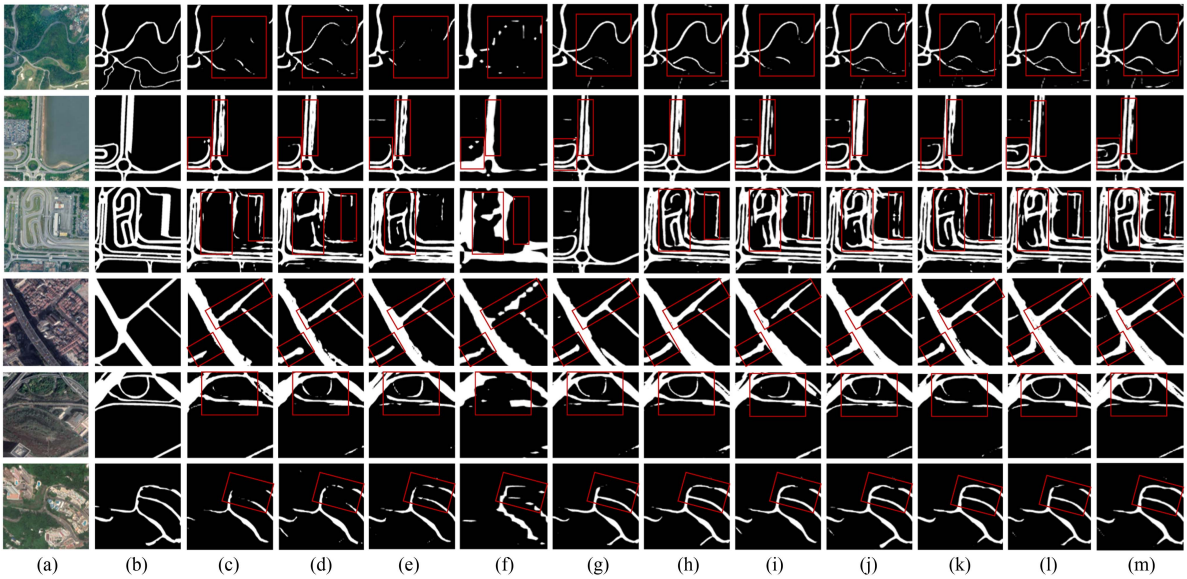

Fig. 9. Comparison of the actual road extraction effect between BIR-Net and mainstream models. The differences are boxed in red. In contrast to other networks, BIR-Net preserves spatial detail and exhibits superior road connectivity and integrity. (a) Original image. (b) Real label. (c) U-Net. (d) Deeplab V3 (ResNet18). (e) PSPNet (MobileNet V2). (f) Mobile ViT. (g) BiseNet V2. (h) DDRNet. (i) D-LinkNet. (j) STDC. (k) PIDNet. (l) PP-LiteSeg. (m) Ours.

the binarized images. Fig. 8(b) represents the real label, and Fig. 8(c)–(h) shows the effect of different techniques on road extraction. Furthermore, the binarized image distinguishes the road and the background by white and black pixels, respectively. The red rectangular box highlights the impact of various techniques on road extraction.

The comparison of Fig. 8(c) and (d) reveals that the proposed deep fusion module is capable of fully leveraging detail branch and semantic branch information to extract a more comprehensive and continuous road region. Additionally, the introduction of ghost convolution (GConv) in Fig. 8(e) enhances the road feature extraction capability, resulting in more comprehensive road extraction. Fig. 8(c) and (f) demonstrates that the proposed RPAM can concentrate on the road's shape in a specific direction and capture the road's distant dependency, resulting in a more continuous road extraction region.

The results of the comparison between Fig. 8(c) and (g)–(i) demonstrate that the SIM is effective in extracting the real roads and distinguishing similar backgrounds. This is achieved by comprehensively utilizing spatial information and contextual information, which effectively addresses the issue of high intraclass variability and interclass similarity in remote sensing images. This enables the extraction of more comprehensive road regions. Ultimately, our BIR-Net can extract more complete and continuous road regions, which is advantageous for urban road planning.

D. Analysis of the Effects of Road Extraction Compared

To validate the road extraction effect of the BIR-Net network, this article tests it on slender and curved urban images and mountainous images obscured by vegetation. Binarized images are used for comparative analysis with other mainstream models. In the images, black pixels represent the background and white pixels represent the roads. The road extraction effect of different models is highlighted by a red rectangular box in Fig. 9.

The comparison of Fig. 9(e) and (f) with Fig. 9(m) reveals that the lightweight models fail to extract the fine curved roads, resulting in incomplete road networks. A comparison of Fig. 9(c), (d), (i), and (m) reveals that some large network models continue to encounter difficulties in extracting missing and discontinuous roads when confronted with slender and curved roads. In contrast, our BIR-Net can effectively address these challenges, resulting in more accurate road extraction.

Based on the analysis of Fig. 9(g), (h), (j), (k), and (l), it is evident that the BiseNet V2, DDRNet, STDC, PIDNet, and PP-LiteSeg networks exhibit a high level of completeness in extracting roads, effectively capturing most road details. However, these networks still encounter some difficulties in accurately extracting roads in areas with vegetation occlusion and pixels with high recognition difficulty, resulting in intermittent road extraction.

Finally, comparing Fig. 9(b) and (m), it is evident that the road region extracted by our BIR-Net has the highest overlap with the actual region. In conclusion, BIR-Net is an effective method for extracting complete and continuous roads from remote sensing images of heavily occluded, slender, and curved roads. This is beneficial for urban road planning as it ensures the use of sufficient lightness.

VI. Conclusion

This article presents a lightweight and efficient urban road network extraction model, BIR-Net, which employs a dual-branch feature extraction backbone network comprising detail branch and semantic branch. An SIM is designed, which effectively addresses the issue of high intraclass variability and high interclass similarity in remote sensing images by synthesizing spatial

and contextual information. A deep fusion module is proposed through a bilateral-guided aggregation approach with the objective of realizing the function of complementary information between detail branches and semantic branches. Furthermore, an RPAM is proposed, which can prioritize the road region and suppresses the background region by adjusting the attention weights and reweighting them. This effectively addresses the issue of intermittent and incomplete extracted road regions due to elongated and curved roads in remote sensing images. The number of parameters and FLOPs of BIR-Net are 3.66 M and 6.49 G, respectively, and its computational efficiency is relatively superior to that of other mainstream models. The road segmentation accuracies of BIR-Net on the CHN6-CUG and DeepGlobe datasets are 59.27% and 58.36%, respectively. These results demonstrate the effectiveness of BIR-Net in extracting roads from remote sensing images. In consideration of the model size and segmentation accuracy, BIR-Net is deemed suitable for deployment on small devices with limited resources.

REFERENCES

[1] Z. Hong, D. Ming, K. Zhou, Y. Guo, and T. Lu, "Road extraction from a high spatial resolution remote sensing image based on richer convolutional features," *IEEE Access*, vol. 6, pp. 46988–47000, Aug. 2018, doi: 10.1109/ACCESS.2018.2867210.

[2] N. Jiang, J. Li, J. Yang, J. Lin, and B. Lu, "A road extraction method of a high-resolution remote sensing image based on multi-feature fusion and the attention mechanism," *Traitement du Signal*, vol. 39, no. 6, pp. 1907–1916, Dec. 2022.

[3] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1826–1848, May 2020, doi: 10.1109/TITS.2019.2913998.

[4] J. Zhang, L. Chen, C. Wang, L. Zhuo, Q. Tian, and X. Liang, "Road recognition from remote sensing imagery using incremental learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 2993–3005, Nov. 2017, doi: 10.1109/TITS.2017.2665658.

[5] J. Mena and J. Malpica, "An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery," *Pattern Recognit. Lett.*, vol. 26, no. 9, pp. 1201–1220, Jul. 2005.

[6] C. Zhu, W. Shi, M. Pesaresi, L. Liu, X. Chen, and B. King, "The recognition of road network from high-resolution satellite remotely sensed data using image morphological characteristics," *Int. J. Remote Sens.*, vol. 26, no. 24, pp. 5493–5508, Apr. 2005.

[7] L. Chen, Q. Zhu, X. Xie, H. Hu, and H. Zeng, "Road extraction from VHR remote-sensing imagery via object segmentation constrained by Gabor features," *ISPRS Int. J. Geoinf.*, vol. 7, no. 9, Sep. 2018, Art. no. 362.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[9] T. Zhang, S. Xiang, W. Liu, Y. Han, X. Guo, and Y. Hao, "Hybrid spiking fully convolutional neural network for semantic segmentation," *Electronics*, vol. 12, Aug. 2023, Art. no. 3565.

[10] H. Zunair and A. Hamza, "Sharp U-Net: Depthwise convolutional network for biomedical image segmentation," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104699.

[11] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, Sep. 2018.

[12] B. Du et al., "Landslide susceptibility prediction based on image semantic segmentation," *Comput. Geosci.*, vol. 155, Oct. 2021, Art. no. 104860.

[13] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Sep. 2021.

[14] M. Fan et al., "Rethinking BiSeNet for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9711–9720.

[15] G. Wang, W. Yang, K. Ning, and J. Peng, "DFC-UNet: A U-net-based method for road extraction from remote sensing images using densely connected features," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, Jun. 2024, doi: 10.1109/LGRS.2023.3329803.

[16] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshop*, Salt Lake City, UT, USA, Jun. 2018, pp. 192–1924, doi: 10.1109/CVPRW.2018.00034.

[17] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, St. Petersburg, FL, USA, Dec. 2017, pp. 1–4, doi: 10.1109/VCIP.2017.8305148.

[18] Q. Zhu et al., "A global context-aware and batch-independent network for road extraction from VHR satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 353–365, May 2021.

[19] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[20] R. Wu, J. Cai, and G. Liu, "Rural road network extraction for high resolution imagery using RDU-net deep learning method," *Remote Sens. Inform.*, vol. 36, no. 1, pp. 29–36, Jan. 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jul. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[22] Y. Li, L. Guo, J. Rao, L. Xu, and S. Jin, "Road segmentation based on hybrid convolutional network for high-resolution visible remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 613–617, Apr. 2019, doi: 10.1109/LGRS.2018.2878771.

[23] A. Filho, M. Shimabukuro, and A. D. Poz, "Deep learning multimodal fusion for road network extraction: Context and contour improvement," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jul. 2023, Art. no. 5001705, doi: 10.1109/LGRS.2023.3291656.

[24] X. Tan, Z. Xiao, Q. Wan, and W. Shao, "Scale sensitive neural network for road segmentation in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 533–537, Mar. 2021, doi: 10.1109/LGRS.2020.2976551.

[25] Y. Lin, F. Jin, D. Wang, S. Wang, and X. Liu, "Dual-task network for road extraction from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 66–78, Jun. 2023, doi: 10.1109/JSTARS.2023.3289217.

[26] S. Zhang, Y. Cao, and B. Sui, "DTHNet: Dual-stream network based on transformer and high-resolution representation for shadow extraction from remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jun. 2023, Art. no. 8000905, doi: 10.1109/LGRS.2023.3290176.

[27] Q. Wu, F. Luo, P. Wu, B. Wang, H. Yang, and Y. Wu, "Automatic road extraction from high-resolution remote sensing images using a method based on densely connected spatial feature-enhanced pyramid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3–17, Jan. 2020, doi: 10.1109/JSTARS.2020.3042816.

[28] S. Liu, M. Li, M. Xu, and Z. Zeng, "An improved lightweight U-net for sea ice lead extraction from multipolarization SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Sep. 2023, Art. no. 2000705, doi: 10.1109/LGRS.2023.3318568.

[29] R. Wang, M. Cai, and Z. Xia, "A lightweight high-resolution RS image road extraction method combining multi-scale and attention mechanism," *IEEE Access*, vol. 11, pp. 108956–108966, Sep. 2023, doi: 10.1109/ACCESS.2023.3313390.

[30] J. Yi, Z. Shen, F. Chen, Y. Zhao, S. Xiao, and W. Zhou, "A lightweight multiscale feature fusion network for remote sensing object counting," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5902113, doi: 10.1109/TGRS.2023.3238185.

[31] D. Liu, J. Zhang, Y. Qi, and Y. Zhang, "A lightweight road detection algorithm based on multiscale convolutional attention network and coupled decoder head," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Apr. 2023, Art. no. 6004605, doi: 10.1109/LGRS.2023.3266054.

[32] X. Chen, Q. Sun, W. Guo, C. Qiu, and A. Yu, "GA-net: A geometry prior assisted neural network for road extraction," *Int. J. Appl. Earth. Observ. Geoinf.*, vol. 114, Nov. 2022, Art. no. 103004.

[33] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 19529–19539, doi: 10.1109/CVPR52729.2023.01871.

[34] S. Zou, F. Xiong, H. Luo, J. Lu, and Y. Qian, "AF-Net: All-scale feature fusion network for road extraction from remote sensing images," in *Proc. Digit. Image Comput., Techn. Appl.*, Gold Coast, QLD, Australia, Nov. 2021, pp. 1–8, doi: 10.1109/DICTA52665.2021.9647235.
[35] L. Dai, G. Zhang, and R. Zhang, "RADANet: Road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5602213, doi: 10.1109/TGRS.2023.3237561.
[36] S. Dong and Z. Chen, "Block multi-dimensional attention for road segmentation in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Dec. 2021, Art. no. 6504505, doi: 10.1109/LGRS.2021.3137551.
[37] C. Wang et al., "Toward accurate and efficient road extraction by leveraging the characteristics of road shapes," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 4404616, doi: 10.1109/TGRS.2023.3284478.
[38] Z. Luo, K. Zhou, Y. Tan, X. Wang, R. Zhu, and L. Zhang, "AD-RoadNet: An auxiliary-decoding road extraction network improving connectivity while preserving multiscale road details," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8049–8062, Jun. 2023, doi: 10.1109/JSTARS.2023.3289583.
[39] Y. Wang et al., "DDU-net: Dual-decoder-U-net for road extraction using high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 4412612, doi: 10.1109/TGRS.2022.3197546.
[40] K. Yang, J. Yi, A. Chen, J. Liu, and W. Chen, "ConDinet++: Full-scale fusion network based on conditional dilated convolution to extract roads from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 8015105, doi: 10.1109/LGRS.2021.3093101.
[41] Z. Ge, Y. Zhao, J. Wang, D. Wang, and Q. Si, "Deep feature-review transmit network of contour-enhanced road extraction from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Mar. 2021, Art. no. 3001805, doi: 10.1109/LGRS.2021.3061764.
[42] Y. Wang et al., "Detecting occluded and dense trees in urban terrestrial views with a high-quality tree detection dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 4707312, doi: 10.1109/TGRS.2022.3184300.
[43] L. Qiu, D. Yu, C. Zhang, and X. Zhang, "A semantics-geometry framework for road extraction from remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Apr. 2023, Art. no. 6004805, doi: 10.1109/LGRS.2023.3268647.
[44] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, Sep. 2019, doi: 10.1109/ACCESS.2019.2940527.
[45] R. Xiao, Y. Wang, and C. Tao, "Fine-grained road scene understanding from aerial images based on semisupervised semantic segmentation networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Dec. 2022, doi: 10.1109/LGRS.2021.3059708.
[46] Y. Li, B. Peng, L. He, K. Fan, and L. Tong, "Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2279–2287, Jul. 2019, doi: 10.1109/JSTARS.2019.2909478.
[47] J. Li et al., "Feature guide network with context aggregation pyramid for remote sensing image segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9900–9912, Nov. 2022, doi: 10.1109/JSTARS.2022.3221860.
[48] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 1577–1586, doi: 10.1109/CVPR42600.2020.00165.
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
[50] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[51] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13708–13717, doi: 10.1109/CVPR46437.2021.01350.
[52] H. Yu et al., "MSAU-net: Road extraction based on multi-headed self-attention mechanism and U-net with high resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Pasadena, CA, USA, Jul. 2023, pp. 6898–6900, doi: 10.1109/IGARSS52108.2023.10281628.
[53] X. Lu et al., "Cascaded multi-task road extraction network for road surface, centerline, and edge extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5621414, doi: 10.1109/TGRS.2022.3165817.
[54] Z. Feng, J. Yang, and Z. Chen, "Urban road network extraction method based on lightweight transformer," *J. Zhejiang Univ.*, vol. 58, no. 1, Jan. 2024, Art. no. 108.
[55] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.
[56] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general purpose, and mobile-friendly vision transformer," Oct. 5, 2021, Accessed: Mar. 4, 2022, *arXiv:2110.02178*.
[57] H. Pan et al., "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023, doi: 10.1109/TITS.2022.3228042.
[58] J. Peng et al., "PP-LiteSeg: A superior real-time semantic segmentation model," Apr. 2022, *arXiv:2204.02681*.

Xianyan Kuang received the Ph.D. degree in traffic information engineering and control from the South China University of Technology, Guangzhou, China, in 2014.

He is currently a Professor with the School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, and Jiangxi Province Key Laboratory of Multidimensional Intelligent Perception and Control, Ganzhou, China. His main research interests include intelligent sensing and control, deep learning and image processing, and intelligent transportation system.

Fujun Cheng is currently working toward the master's degree in artificial intelligence with the School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou, China.

His main research interests include deep learning and image processing, semantic segmentation, and remote-sensing image process.

Cuiqin Wu received the M.S. degree in mechanical and electronic engineering from the Jiangxi University of Science and Technology, Ganzhou, China, in 2006.

Her main research interests include intelligent perception and robotics, and deep learning and image processing. She is currently a Lecturer with the School of Mechanical and Electrical Engineering, Jiangxi University of Science and Technology, Ganzhou, China.

Hui Lei is currently working toward the master's degree in artificial intelligence with the School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou, China.

Her main research interests include deep learning and image processing, and intelligent transportation system.

Zuliang Zhang is currently working toward the master's degree in control engineering with the School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou, China.

His main research interests include deep learning and image processing, and intelligent transportation system.