

Road Extraction from High-Resolution Remote Sensing Images Based on MRCBL-Net and Multi-attention Mechanism

Jiajia Liu ^{1,*}, Xuan Zhao ², Wenxiang Dong ², Chidong Huang ², Jianhua Liu ², Xiaoguang Tu ²

¹ Civil Aviation Flight University of China, Center for Faculty Development and Teaching Evaluation, Guanghan, Deyang, Sichuan, 618307, China

² Civil Aviation Flight University of China, College of Aviation Electronic and Electrical Engineering, Guanghan, Deyang, Sichuan, 618307, China

* Corresponding author: Jiajia Liu (Email: cafucljj@163.com)

Abstract: Road extraction from high-resolution remote sensing images is crucial for urban planning and geographic information systems (GIS). However, complex background interference, severe occlusions, and the inherent morphological complexity of roads often lead to discontinuities and insufficient accuracy in extraction results. To address these challenges, this paper proposes a Multi-scale Refined Convolutional Backbone Link Network (MRCBL-Net). The network adopts ConvNeXt as the backbone encoder to enhance feature extraction capability and embeds the proposed Multi-scale Dilated Convolution and Spatial-Channel Refinement (MULCSR) module into skip connections. The core of this module is the Channel and Spatial Relation (CSR) module, which captures multi-scale contextual information through parallel multi-branch dilated convolutions, and adaptively calibrates feature weights via channel attention and spatial attention mechanisms in sequence. This focuses on key road features, thereby effectively fusing shallow details and deep semantics. Experiments on the DeepGlobe dataset demonstrate that the proposed method achieves an accuracy of 98.61%, precision of 82.98%, recall of 82.46%, F1-score of 82.52%, and Intersection over Union (IoU) of 70.55%. Experiments on the CHN6-CUG dataset show that the method attains an accuracy of 96.13%, precision of 79.79%, recall of 79.66%, F1-score of 79.28%, and IoU of 66.16%. Compared with existing methods, the proposed method achieves the optimal performance in all five metrics. Qualitative results further confirm that the proposed method excels in maintaining road continuity and restoring fine boundaries, improving the robustness and accuracy of road extraction in complex scenarios.

Keywords: Road Extraction; Semantic Segmentation; High-resolution Remote Sensing Images; Multi-scale Feature Fusion.

1. Introduction

With the continuous development of remote sensing technology and the widespread application of remote sensing image data, road extraction from remote sensing images has become a key research task in geographic information systems (GIS) [1]. As an important component of urban infrastructure, the significance of roads in areas such as traffic management, urban planning, post-disaster assessment, and environmental monitoring is self-evident [2]. Therefore, how to accurately extract road information from complex remote sensing images has become a core challenge in the field of remote sensing image processing [3].

To address the aforementioned challenges, various deep learning-based methods for road extraction have emerged in the field of semantic segmentation. U-Net[4], with its U-shaped encoder-decoder architecture and skip connection mechanism, achieves precise pixel-level prediction through symmetric upsampling and downsampling paths. Architectures based on U-Net have derived a series of road extraction networks. For instance, DeepLabV3+[5] introduces a decoder module on the basis of DeepLab, maintains pixel spatial relationships through dilated convolutions, and effectively aggregates multi-scale road features using Atrous Spatial Pyramid Pooling (ASPP). D-LinkNet [6] adds a center module between the encoder and decoder, employing cascaded dilated convolutions to expand the receptive field, enabling it to utilize low-level spatial

details to refine road boundaries while extracting high-level semantic features. Although these CNN-based models have achieved results in road extraction tasks, purely CNN-based models (such as DeepLabV3+, D-LinkNet) suffer from inherent limitations in modeling long-range global dependencies due to their local receptive fields, making it difficult to effectively handle road discontinuities caused by occlusion or shadows. In recent years, the introduction of the Transformer architecture has brought new ideas. Vision Transformer (ViT) [7] applies the pure Transformer architecture to image classification. By dividing images into patches, introducing positional encodings, and utilizing self-attention mechanisms, it effectively captures global contextual information. SEG-Road[8] proposes a hybrid architecture combining Transformer and CNN, leveraging Transformer to capture global dependencies while preserving local details through CNN, and introduces connectivity structures to optimize road topology. Zhu et al. [9] constructed a spatial attention mechanism based on Swin Transformer, effectively improving the accuracy of road extraction. However, Transformer-based models or hybrid models (e.g., SEG-Road, Swin Transformer method), while improving global contextual modeling capability, often require significant computational overhead and memory consumption, limiting practical application efficiency on high-resolution remote sensing images. To address this issue, A2-FPN [10] adopts an adaptive feature pyramid network to fuse multi-scale features. DSNet[11] uses a novel application of dilated convolutions aimed at more efficiently expanding

the receptive field. Although these studies have made progress at different levels, their multi-scale fusion strategies mostly still lack adaptive attention mechanisms for channel and spatial dimensions, failing to achieve adaptive weighting and interaction between information at different scales, which limits the generalization ability for roads of varying widths and morphologies in complex scenes and makes it difficult to focus on key features and suppress complex background interference.

To improve the model's generalization capability and effectively address the accuracy problem of road extraction, a novel multi-attention fusion module named MULCSR is proposed. This module is embedded into skip connections, constructing a shallow content-guided attention mechanism, the core of which is to use shallow features rich in details to guide the optimization of deep semantic features. The main contributions are as follows:

- (1) A parallel multi-scale feature extraction and fusion architecture (MULCSR) is proposed, achieving cross-scale contextual perception through a parallel multi-branch dilated convolution structure. This framework enables the network to adaptively select and fuse features based on the input content, enhancing generalization ability for roads of different morphologies.
- (2) A sequentially processed CSR module is proposed, which realizes recalibration of feature channels through channel attention, combined with spatial attention to achieve precise focusing on key regions, forming a hierarchical feature optimization mechanism.
- (3) A complete optimization pipeline of "detail guidance-context perception-feature focusing" is constructed. Through

the interaction of three levels—shallow feature guidance, multi-scale parallel processing, and channel-spatial synergy—the boundary accuracy and structural continuity of road extraction are comprehensively improved.

(4) Experimental results on the DeepGlobe dataset show that the proposed method performs excellently in the completeness of road extraction, achieving an accuracy of 98.61%, a precision of 82.98%, a recall of 82.46%, an F1-score of 82.52%, and an Intersection over Union (IoU) of 70.55%. Experiments on the CHN6-CUG dataset show that the proposed method achieves an accuracy of 96.13%, a precision of 79.79%, a recall of 79.66%, an F1-score of 79.28%, and an IoU of 66.16%.

2. Method

2.1. Overall Network Architecture

To improve the continuity and precision of road extraction in high-resolution remote sensing images, a novel multi-attention combined MRCBL-Net network is proposed. The core architecture of the MRCBL-Net network consists of a ConvNeXt backbone network, attention modules, the MULCSR module, and a decoder structure based on D-LinkNet. The core of this network is to introduce the Multi-scale Dilated Convolution and Spatial-Channel Refinement module (MULCSR) into the skip connections between the encoder and decoder, thereby constructing a shallow detail-guided attention mechanism to achieve adaptive fusion of multi-scale contextual information. The MRCBL-Net network structure is shown in Figure 1.

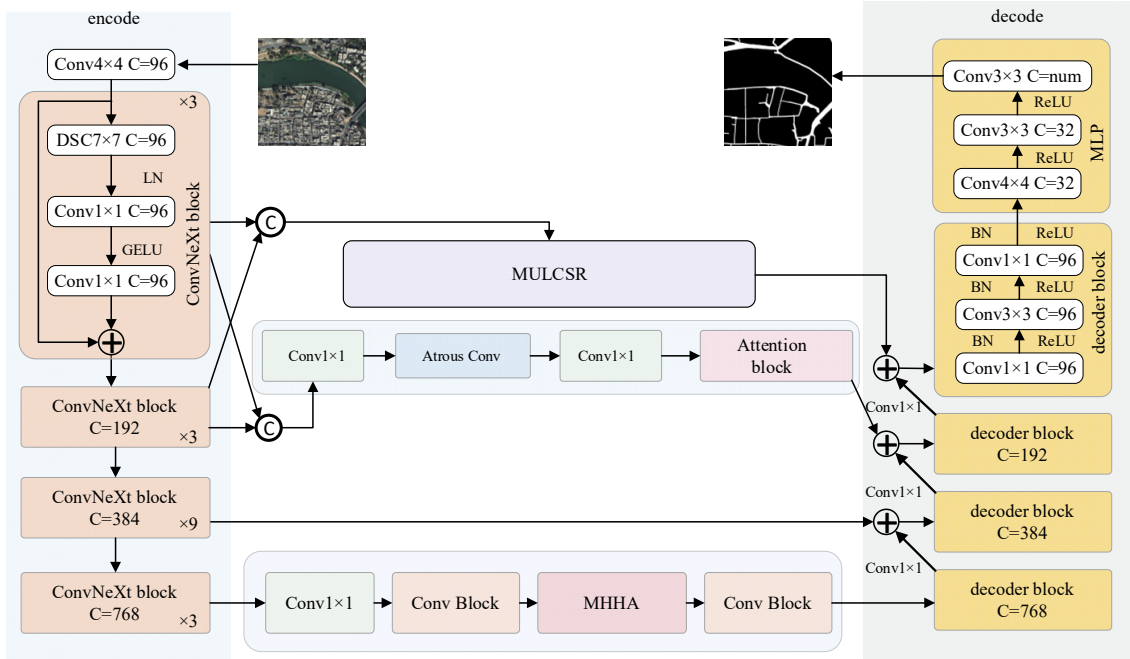


Fig 1. MRCBL-Net Network Structure

The ConvNeXt encoder is the core feature extraction part of the network. This structure ensures high feature extraction capability while maintaining high computational efficiency. The MRCBL-Net network initially processes the image for feature extraction through the ConvNeXt encoder, outputting features at four different hierarchical levels (shallow, lower-middle, middle, and deep). Each ConvNeXt block is stacked with residual connections, further enhancing the network's expressive power and improving the stability of the deep

network. Shallow features, rich in spatial details (such as edges, textures), are processed by the MULCSR module to achieve adaptive selection at different scales, focusing on balancing road detail preservation and context capture. The middle-level feature maps employ a lightweight attention mechanism to extract detailed features. Deep features contain high-level semantic information of the image. They are first adjusted in the number of channels through a 1×1 convolutional layer, then fed into the Multi-Head Hybrid

Attention module (MHHA) to help the model focus on different parts of the input, thereby efficiently capturing global dependencies and improving the understanding of the overall road structure.

2.2. MULCSR Module

The core idea of this module is to capture multi-scale

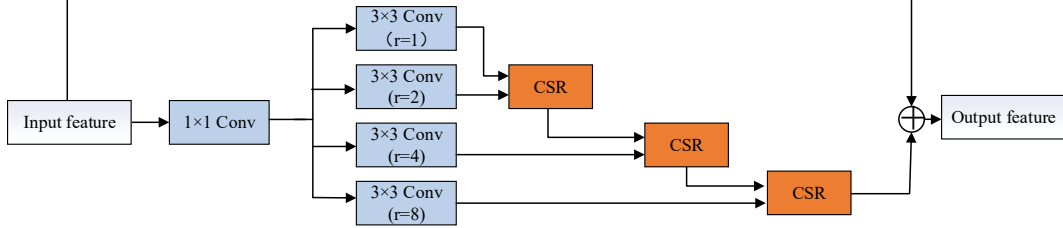


Fig 2. MULCSR Module Structure

The input feature map F_{in} is first passed through a 1×1 convolutional layer for channel dimension adjustment, and then sent to four parallel branches. Each branch uses a 3×3 dilated convolutional layer, with dilation rates (r) set to 1, 2, 4, and 8, respectively. By setting different dilation rates, each branch can possess a differentiated receptive field, thereby focusing on local details and global context, respectively. The output feature map F_r of the r -th branch can be expressed as:

$$F_r = \text{Conv}_{3 \times 3}^{d=r}(F_{in}) \quad (1)$$

where $\text{Conv}_{3 \times 3}^{d=r}$ represents the 3×3 dilated convolution operation with dilation rate r .

Different from traditional methods that typically directly concatenate or sum multi-scale features, MULCSR sends the features F_r extracted by each branch into an independent CSR module for refinement. This process can be expressed as:

$$F_r^{CSR} = \text{CSR}(F_r) \quad (2)$$

In Equation (2), $\text{CSR}(\cdot)$ represents a multi-attention module.

features through a parallel multi-branch structure and refine the features at each scale using the designed CSR (Channel and Spatial Relation) module, ultimately achieving feature fusion. Its structure is shown in Figure 2.

All refined multi-scale feature maps $\{F_r^{CSR} | r = 1, 2, 4, 8\}$ are finally fused by summation along the channel dimension to obtain the final output of the module F_{MULCSR}^{CSR} :

$$F_{MULCSR} = \sum_{r \in \{1, 2, 4, 8\}} F_r^{CSR} \quad (3)$$

This fusion strategy ensures that the integrated features can adaptively highlight the context information most relevant to the current road scene.

2.3. CSR Module

This module introduces a coordinated attention mechanism in channel and spatial dimensions, achieving multi-level refined processing of feature maps. The overall architecture of the CSR module adopts a sequential processing flow, establishing a hierarchical attention mechanism. This design ensures that the calculation of spatial attention is based on features optimized in the channel dimension, thereby achieving precise focus on key features. The overall architecture of the CSR module is shown in Figure 3.

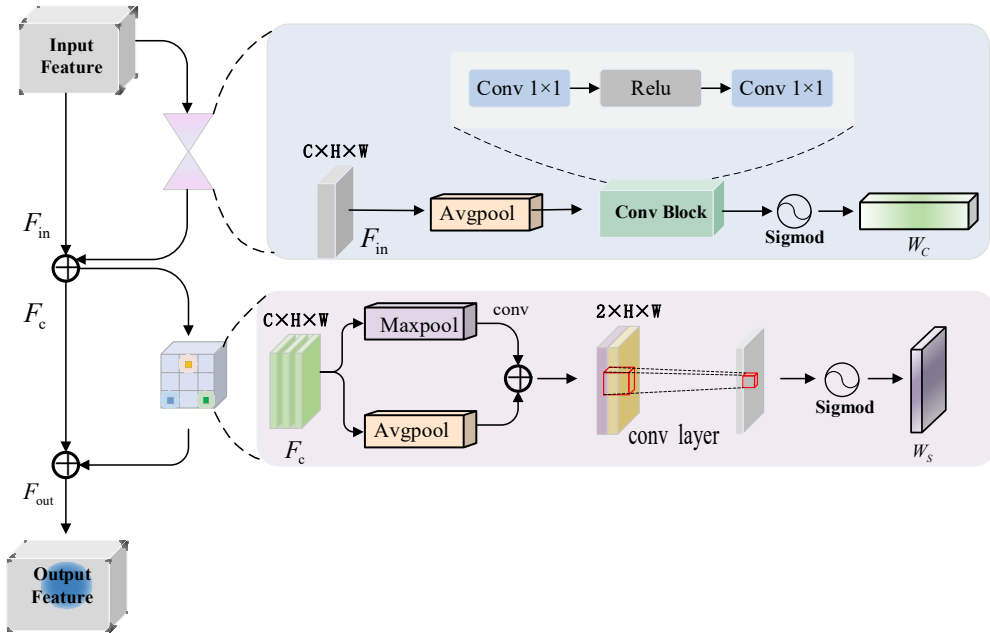


Fig 3. CSR Module Structure

The input to the CSR module is the feature map extracted by the previous convolutional layer. These feature maps carry low-dimensional representations of key information in the image.

The input feature map first undergoes a 1×1 convolution to adjust the number of channels. The purpose of this operation is to reduce the channel count of the feature map, improve

computational efficiency, and simultaneously provide more precise channel information for subsequent convolutional layers. The convolved feature map undergoes a nonlinear transformation via the ReLU activation function, further enhancing the network's expressive power. The feature map after ReLU activation is again subjected to feature mapping through a 1×1 convolution, and the Sigmoid activation function is applied. Sigmoid compresses the value of each channel to between 0 and 1, representing the importance of each channel. This calculation process is implemented through sequential convolution and activation operations:

$$\mathbf{W}_c = \sigma(\text{Conv}_{1 \times 1}(\delta(\text{Conv}_{1 \times 1}(\mathbf{F}_{in})))) \quad (4)$$

where $\mathbf{F}_{in} \in \mathbb{R}^{C \times H \times W}$ is the input feature map, $\text{Conv}_{1 \times 1}$ represents the 1×1 convolution operation. δ is the ReLU activation function, and σ is the Sigmoid activation function.

The obtained channel attention weight $\mathbf{W}_c \in \mathbb{R}^{C \times 1 \times 1}$ is multiplied element-wise with the input feature along the channel dimension, completing the feature optimization in the channel dimension:

$$\mathbf{F}_c = \mathbf{W}_c \cdot \mathbf{F}_{in} \quad (5)$$

This step enables the network to automatically enhance channels related to road features and suppress irrelevant background channel interference.

After Sigmoid activation, the feature map is processed through spatial pooling (average pooling and max pooling) to extract spatial features. The feature maps after pooling and convolution processing are summed to obtain the final spatial attention feature map. Finally, the output feature map is multiplied element-wise with the original input feature map to obtain the final output feature map weighted by spatial attention. The calculation process is expressed as:

$$\mathbf{W}_s = \sigma(\text{Conv}_{1 \times 1}([\text{AvgPool}(\mathbf{F}_c); \text{MaxPool}(\mathbf{F}_c)])) \quad (6)$$

where, AvgPool represents Global Average Pooling along the channel dimension, with an output size of $1 \times H \times W$. MaxPool represents Global Max Pooling along the channel dimension, with an output size of $1 \times H \times W$. The symbol $[:]$ in Equation (6) represents feature concatenation, and the concatenated feature size is $2 \times H \times W$.

The final output feature map $\mathbf{F}_{out}$ is:

$$\mathbf{F}_{out} = \mathbf{W}_s \cdot \mathbf{F}_c \quad (7)$$

Through this sequential operation, the CSR module can dynamically enhance the response to discriminative feature channels and key spatial locations of roads.

Through multiple convolution, pooling, and weighting operations, the CSR module ensures effective fusion of multi-scale information, avoiding information loss. The CSR module enhances the network's performance in the road extraction task from remote sensing images. The network can not only extract global features but also capture local details of roads, improving the model's robustness and accuracy.

2.4. Loss Function

This paper adopts a hybrid loss function that combines Binary Cross-Entropy (BCE) loss [12] and Dice loss [13], aiming to enhance the model's segmentation performance. BCE loss captures structural information to preserve image texture, while Dice loss emphasizes edge details in high-frequency regions. This combination enhances the learning capability for the road target area, improves the network's adaptability to imbalanced classes, and ensures the integrity of global features while extracting fine edge features. The specific expression of this loss function is as follows:

$$L_{BCE} = 1 - \frac{2 \sum_i^{H \times W} \phi_{true}^i \phi_{pred}^i + 1}{\sum_i^{H \times W} (\phi_{true}^i)^2 + \sum_i^{H \times W} (\phi_{pred}^i)^2 + 1} \quad (8)$$

$$L_{Dice} = -\frac{1}{N} \sum_{i=1}^N (\gamma \phi_{true}^i \log(\phi_{pred}^i))$$

$$+ (1 - \phi_{true}^i) \log(1 - \phi_{pred}^i) \quad (9)$$

$$L_{Hybrid} = L_{BCE}(\delta_{true}, \delta_{pred}) + L_{Dice}(\delta_{true}, \delta_{pred}) \quad (10)$$

where N is the number of pixels in the sample. The binary class label ϕ_{true}^i for pixel i represents the ground truth label of the sample, taking a value of 0 or 1. The predicted probability ϕ_{pred}^i for the sample classification is between 0 and 1. Through the loss functions L_{Road} and L_{Edge} (targeting edge details and sample labels, respectively) and the prediction result loss functions δ_{true} and δ_{pred} (targeting overall road texture), the accuracy of the segmentation model and edge information extraction is improved to varying degrees.

3. Experiments

3.1. Evaluation Metrics

The road extraction task from remote sensing images is typically measured by five evaluation metrics: Precision, Recall, Intersection over Union (IoU), F1-score, and Accuracy [14]. Specifically, Precision represents the proportion of correctly predicted road pixels among all pixels predicted as road, calculated as:

$$\text{Pre} = \frac{TP}{TP + FP} \quad (11)$$

Recall measures the proportion of correctly predicted road pixels to the total number of actual road pixels in the ground truth image. Its calculation formula is:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (12)$$

Accuracy refers to the percentage of correctly predicted pixels among the total predicted pixels. Its calculation formula is:

$$\text{Acc} = \frac{TP + TN}{TP + FN + FP + TN} \quad (13)$$

F1-score is the harmonic mean of Precision and Recall. Its calculation formula is:

$$F_1 = \frac{2 \times \text{Pre} \times \text{Recall}}{\text{Pre} + \text{Recall}} = \frac{2TP}{2TP + FN + FP} \quad (14)$$

Intersection over Union reflects the overlap ratio between predicted road pixels and ground truth pixels. Its calculation formula is:

$$\text{IoU} = \frac{TP}{TP + TN + FP} \quad (15)$$

3.2. Experimental Datasets and Setup

All experiments were conducted under the same environment and the same parameters to ensure fair comparison. Computations were performed on a Windows 11 system using one NVIDIA GeForce RTX 4070 (12GB GDDR6) GPU, with a CPU of Hygon C86-3G (OPN:3350). The programming language was Python 3.12, and PyTorch 2.5.1 + CUDA 12.1 was used as the deep learning framework.

Experiments in this paper were conducted on the DeepGlobe dataset and the CHN6-CUG road dataset. The DeepGlobe dataset contains remote sensing images of different scenes, each with annotation information for road areas. The image resolution is 1024×1024 pixels, with a

spatial resolution of 0.5 meters/pixel. The dataset is divided into 4980 training images, 250 validation images, and 996 test images for experiments. The CHN6-CUG road dataset focuses on typical urban agglomeration regions in China. Each image has a size of 512×512pixels, with a spatial resolution of 0.5 meters/pixel. The images in the dataset are divided into 3608 training images and 903 test images for experiments. During the training phase, various data augmentation techniques were applied to the input images, including random color jitter, random rotation, translation, horizontal and vertical flipping, etc., to enhance the network's robustness and improve the model's generalization ability.

The Adam optimizer was used in the experiments, with the batch size set to 4 and the initial learning rate set to 2×10^{-4} . To achieve an efficient training process, a multi-step learning

rate scheduling strategy was adopted: if the loss value did not show a significant decrease for 6 consecutive training epochs, the learning rate was halved; training automatically terminated when the learning rate dropped below 5×10^{-8} . To further improve the accuracy of road extraction and accelerate model convergence, all network backbones were initialized with weights pre-trained on the ImageNet-1K dataset.

3.3. Ablation Experiments

To systematically verify the impact of the backbone network and the MULCSR module on network performance, ablation experiments were conducted, and the results are shown in Table 1.

Table 1. Results of Ablation Experiments

Group	Backbone	MULCSR	F1-score (%)	IoU (%)
1	Resnet 50		71.89	58.02
2	Resnet 50	√	72.13	61.37
3	ConvNext		76.15	62.97
4	ConvNext	√	82.52	70.55

From Table 1, it can be seen that the model with ResNet-50 as the backbone and without the MULCSR module achieves an F1-score and IoU of 71.89% and 58.02%, respectively. After introducing the MULCSR module, the model's IoU metric reaches 61.37%, an improvement of 3.35%. This improvement verifies the effectiveness of the MULCSR module itself. This module enhances the network's ability to capture contextual information through its multi-scale dilated convolution and attention refinement mechanism, and optimizes feature representation using the CSR module, thereby compensating for the shortcomings of a single backbone network in feature extraction. Replacing the backbone network with ConvNeXt without introducing the MULCSR module yields an F1-score and IoU of 76.15% and 62.97%, respectively. This demonstrates the good feature

representation capability of the ConvNeXt backbone network, whose architecture can extract richer and more discriminative basic features from the input image. After introducing the MULCSR module, the F1-score and IoU metrics reach 82.52% and 70.55%, respectively. Compared with group 3, IoU improves by 6.81%. This indicates a synergistic enhancement effect between the ConvNeXt backbone and the MULCSR module. The backbone network provides high-quality initial features, while the MULCSR module performs precise multi-scale context perception and feature refinement based on this. Their combination jointly optimizes feature representation and improves the accuracy of road extraction. For qualitative analysis, Figure 4 shows the visualization comparison results of groups 1-4 on the DeepGlobe road dataset.

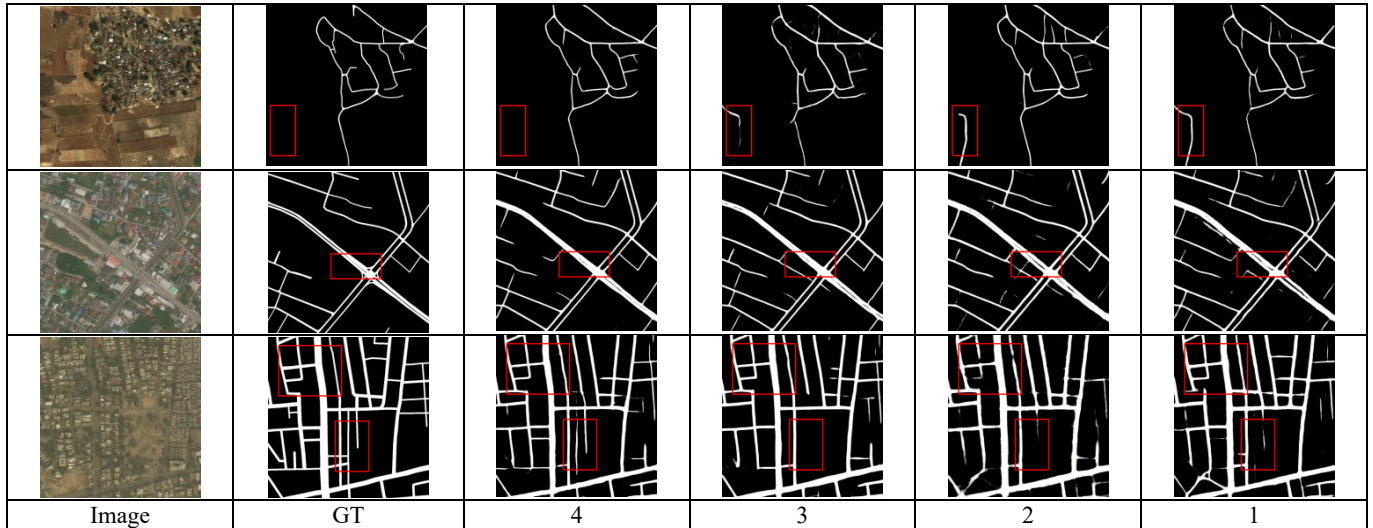


Fig 4. Visual Comparison of Ablation Experiments

The qualitative visualization results in Figure 4 show the performance differences of different models in complex scenes. Models in groups 1 and 2 are susceptible to interference, producing more false detections or missed detections. The model in group 3 shows improved anti-interference capability. The model in group 4 can most effectively suppress background noise and accurately focus

on road areas. This reflects the robustness of the MULCSR module in complex scenes.

The results of the ablation experiments show that the architecture with ConvNeXt as the backbone and embedded with the MULCSR module outperforms the baseline model on all metrics, verifying the rationality of the core module design.

3.4. Loss Function Comparative Analysis

To verify the effectiveness of the hybrid loss function used in this paper, the performance differences among using only BCE loss, only Dice loss, and the combination of the two (hybrid loss) under the same experimental configuration are

compared. Through quantitative metrics and qualitative visualization comparisons, the impact of different loss functions on road extraction results is deeply analyzed, especially regarding the preservation of edge details and structural continuity.

Table 2. Comparison Results of Loss Functions on the DeepGlobe Road Dataset

Loss Function	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
BCE	98.35	82.22	79.95	80.91	68.26
Dice	98.27	82.10	82.17	81.95	68.79
BCE+Dice	98.61	82.98	82.46	82.52	70.55

From Table 2, it can be seen that the hybrid loss function achieves the best results on all five metrics. When using only BCE loss, the model achieves an accuracy of 98.35%, but its recall is only 79.95%, and the IoU is 68.26%. This indicates that the model pays more attention to pixel-level classification accuracy but performs inadequately in terms of road edge continuity. When using only Dice loss, the model's recall increases to 82.17%, providing some alleviation for the class imbalance problem, but the false detection rate under

complex backgrounds is higher. The hybrid loss function combines the advantages of both, achieving a precision of 82.98%, a recall of 82.46%, and an IoU metric of 70.55%. It improves recall and IoU while maintaining high precision, proving that it can better balance the completeness of road areas and boundary accuracy.

To demonstrate the effect differences of different loss functions, Figure 5 provides a visual comparison on the DeepGlobe road dataset.

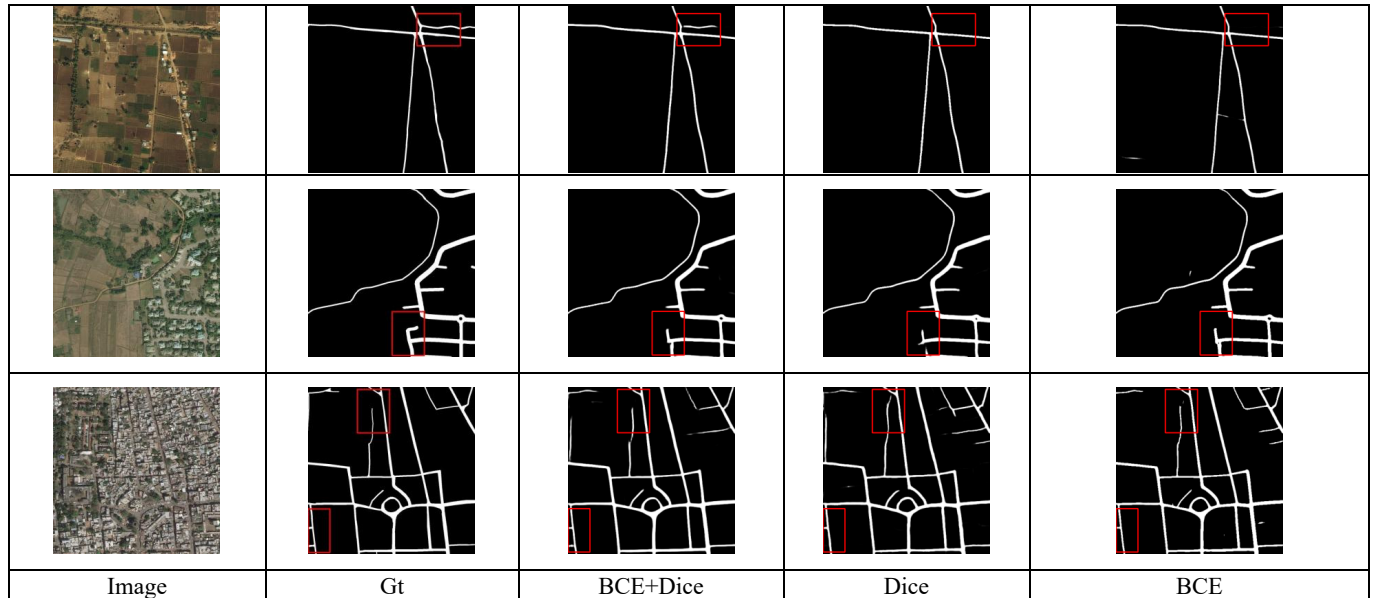


Fig 5. Visual Comparison of Loss Functions on the DeepGlobe Road Dataset

The visualization results in Figure 5 reflect the impact of different loss functions on road extraction results. The model output using only the BCE loss function can identify the main structure of the road, but its predicted structure has structural defects, and there are problems of local missed detection and false detection in some areas (as shown in the red boxes in Figure 5). The per-pixel optimization mechanism of the BCE loss function can guarantee overall classification accuracy but has limitations in understanding the structure of morphologically complex roads. The model output using the Dice loss function shows improved connectivity, and the false detection problem is alleviated. However, this model still suffers from missed detection phenomena. The Dice loss function still has insufficient sensitivity in foreground-background classification. The model with the hybrid loss function in this paper achieves superior visualization results. The extracted road network not only maintains good structural continuity and integrity but also has clearer and

smoother boundaries. The hybrid loss function effectively suppresses background interference while improving road boundary precision, outperforming the performance of a single loss function.

3.5. Comparative Experiments

To comprehensively evaluate the performance of the proposed method in the remote sensing road extraction task, this paper compares the model with various representative semantic segmentation networks. The evaluation uses five commonly used metrics: overall Accuracy, Precision, Recall, F1-score, and Intersection over Union (IoU).

3.5.1. DeepGlobe Road Dataset

To objectively evaluate the performance of the proposed method, Table 3 lists the comparative results of MRCBL-Net on the five core metrics on the DeepGlobe dataset.

Table 3. Comparison Results on the DeepGlobe Road Dataset

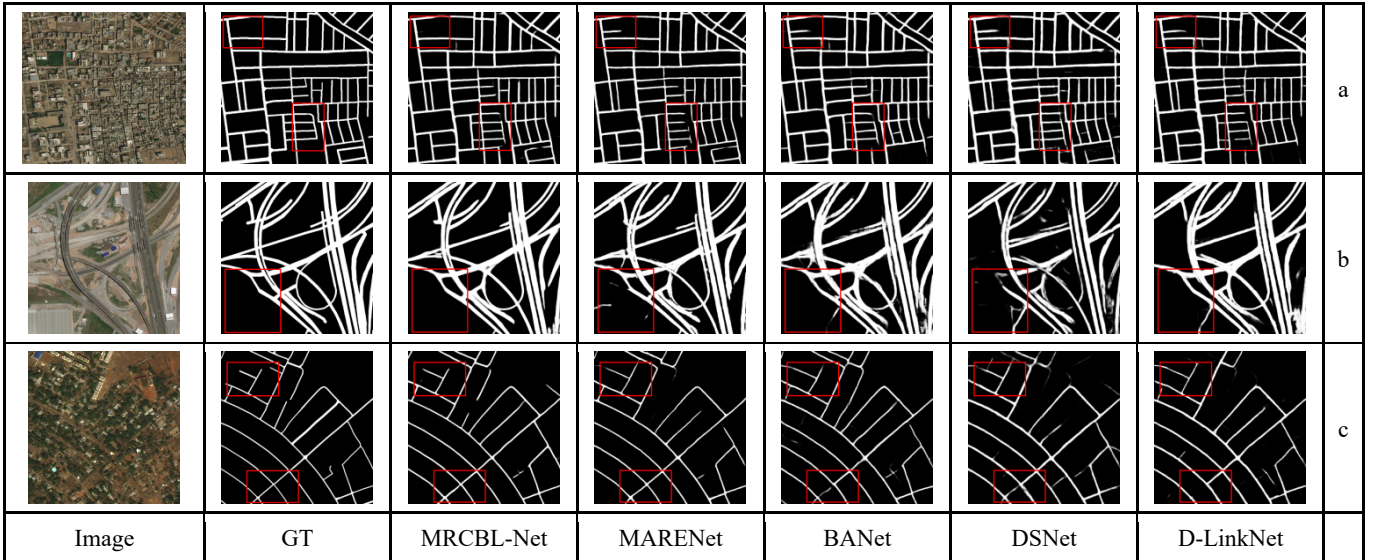
Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
A2FPN [10]	98.01	75.77	76.01	75.89	61.05
DeepLabv3+ [5]	97.98	76.06	73.29	73.79	59.29
D-LinkNet [6]	98.16	77.73	76.48	76.78	62.62
DSNet[11]	98.08	76.28	75.62	75.59	61.17
BASNet[15]	98.34	80.03	78.49	79.01	65.63
FFNet[16]	98.04	75.92	75.68	75.79	60.80
UNet3+[17]	97.90	77.68	72.35	73.08	59.62
UNetformer[18]	97.97	73.51	78.14	75.75	60.95
MARENet[19]	98.49	81.39	80.77	80.69	68.22
improve Deeplab3+[20]	97.89	79.65	81.66	80.66	67.35
MRCBL-Net	98.61	82.98	82.46	82.52	70.55

From the quantitative comparison results in Table 3, it can be seen that MRCBL-Net achieves the highest accuracy among all compared models, reaching 98.61%, demonstrating the model's robustness on large-scale remote sensing images. MRCBL-Net achieves precision and recall of 82.98% and 82.46%, respectively. The precision is 1.59% higher than that of the MARENet model, indicating that MRCBL-Net has a greater advantage in suppressing background false detections. The recall is 0.55% higher than that of the improve-U-net model. Improving recall without sacrificing precision indicates that MRCBL-Net provides more sufficient identification for weakly textured, occluded, and disconnected road segments. MRCBL-Net achieves an F1-score of 82.52%, which is 1.83% higher than that of MARENet. This metric synthesizes the trade-off between precision and recall, further proving the stability and precision of MRCBL-Net's segmentation results. MRCBL-Net achieves an IoU of 70.55%, which is 2.33% higher than

that of MARENet. The advantage in IoU intuitively reflects the improvement in spatial overlap and boundary consistency, especially beneficial for the boundary delineation of thin roads and complex backgrounds.

Comprehensive evaluation results across the five metrics show that MRCBL-Net exhibits the best performance among all compared models, verifying the effectiveness and generalization ability of the proposed architecture in the remote sensing road segmentation task. The model not only achieves improvements in various quantitative metrics but also demonstrates superior detail reconstruction capability in complex scenarios such as complex backgrounds, weak textures, and slender objects.

To further verify the above quantitative analysis conclusions from a visual dimension, Figure 6 shows the visual comparison of MRCBL-Net on the DeepGlobe road dataset.

**Fig 6.** Visual Comparison of Various Methods on the DeepGlobe Road Dataset

As shown in Fig. 6, MRCBL-Net consistently yields segmentation results closest to the ground truth across all three subfigures. In subfigure (a), it correctly extracts roads in the upper-left region without false detections, whereas MARENet, BANet, DSNet, and D-LinkNet all exhibit erroneous extraction. In subfigures (b) and (c), MRCBL-Net

fully preserves road details and continuity in the lower-left regions, while the compared methods show varying degrees of false detection, discontinuities, and omissions, resulting in loss of road structural information. These findings confirm that MRCBL-Net outperforms traditional methods in structural preservation, particularly in occluded or detail-

blurred areas, thereby enhancing extraction reliability.

3.5.2. CHN6-CUG Road Dataset

The experimental results based on the CHN6-CUG road dataset are shown in Table 4.

Table 4. Comparison Results on the CHN6-CUG Road Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
D-LinkNet[6]	96.01	78.53	77.47	77.64	64.02
DSNet[11]	95.48	76.43	74.70	74.96	60.52
BASNet [15]	95.64	75.97	74.97	74.87	60.72
FFNet[16]	95.40	78.39	67.87	72.16	57.24
UNetformer[18]	95.91	78.73	76.78	76.25	63.50
MARENet [19]	95.44	78.14	78.15	78.12	64.71
MRCBL-Net	96.13	79.79	79.66	79.28	66.16

From Table 4, it can be seen that MRCBL-Net achieves an accuracy of 96.13%, which is 0.12% higher than that of D-LinkNet, demonstrating the model's strong generalization ability across different datasets. MRCBL-Net achieves a precision of 79.79%, which is 1.65% higher than that of MARENet, indicating that the model has good anti-interference ability in complex backgrounds and can effectively suppress false detections in non-road areas. The recall metric is 79.66%, which is 1.51% higher than that of MARENet, indicating that the model's coverage of real road areas is more complete, maintaining high detection sensitivity

especially in weak-texture and partially occluded areas. The F1-score of MRCBL-Net reaches 79.28%, which is 1.16% higher than that of MARENet. MRCBL-Net's Intersection over Union (IoU) metric reaches 66.16%, which is 1.45% higher than that of MARENet, intuitively reflecting the advantage in spatial overlap between prediction results and real road areas.

For qualitative analysis, Figure 7 shows the visual comparison results of various methods on the CHN6-CUG road dataset.

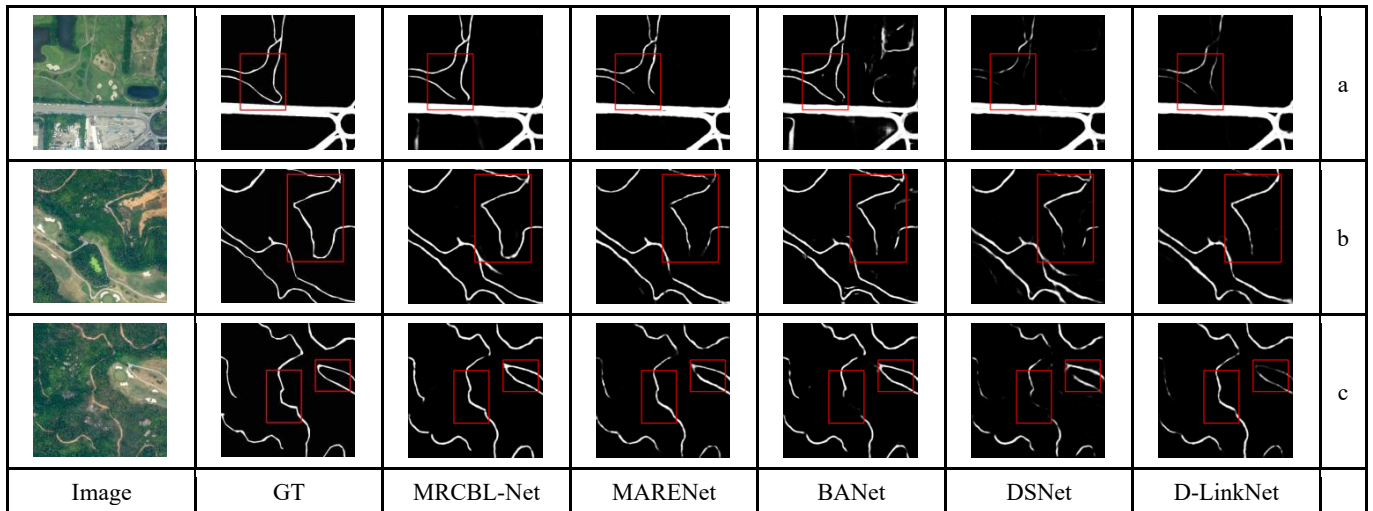


Fig 7. Visual Comparison of Various Methods on the CHN6-CUG Road Dataset

From Figure 7, it can be seen that MRCBL-Net maintains good road continuity compared to other methods. In areas occluded by trees, MRCBL-Net can still effectively identify partially occluded road segments, demonstrating the model's robustness to complex lighting conditions. MRCBL-Net can effectively restore the topological structure of roads, while comparison methods show obvious fractures and boundary blurring problems at curves.

Comprehensive analysis of quantitative and qualitative results shows that MRCBL-Net can maintain stable road extraction performance on remote sensing images of different scenes, especially when dealing with challenges such as tree occlusion and road curvature in complex urban scenarios, MRCBL-Net demonstrates technical advantages.

4. Conclusion

This paper addresses challenges such as complex background interference, easy loss of road details, and difficulty in maintaining continuity in road extraction from high-resolution remote sensing images, proposing a multi-attention fusion-based MRCBL-Net network. The network uses ConvNeXt as the backbone encoder and the D-LinkNet decoder as the basic framework. By introducing our innovatively designed MULCSR module and CSR attention mechanism, the model's ability to capture multi-scale contextual information and focus on key channel and spatial features is enhanced. Comprehensive experiments on the public DeepGlobe and CHN6-CUG datasets show that the proposed MRCBL-Net model achieves superior levels in five core evaluation metrics: Accuracy, Precision, Recall, F1-

score, and Intersection over Union (IoU). Qualitative analysis results further verify that, compared to other mainstream segmentation networks, this model can more completely restore the topological structure of roads and better maintain the fineness of road edges and the continuity of regions when processing complex scenes. MRCBL-Net effectively improves the accuracy and robustness of road extraction from high-resolution remote sensing images. The proposed MRCBL-Net model performs excellently on multiple metrics, showing significant advantages especially in maintaining road continuity and fine boundaries. However, there is still room for improvement in the model's computational efficiency when processing large-scale regions and its generalization ability for extremely narrow roads (such as rural paths) or roads with abnormally complex topology (such as large interchanges). Future work will focus on optimizing the model's computational complexity and exploring its application to more challenging remote sensing image interpretation tasks, such as multi-temporal road change detection.

References

- [1] Miao, Z., Fu, K., Yan, M., & et al. (2023). Road extraction from remote sensing imagery: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 333-352. <https://doi.org/10.1016/j.isprsjprs.2022.12.012>.
- [2] Liu, R., Wu, J., Lu, Q., & et al. (2024). A review of deep learning-based methods for road extraction from high-resolution remote sensing images. *Remote Sensing*, 16(12), 2056. <https://doi.org/10.3390/rs16122056>.
- [3] Wang, X., Jin, X., Dai, Z., & et al. (2025). Deep learning-based methods for road extraction from remote sensing images: A vision, survey, and future directions. *IEEE Geoscience and Remote Sensing Magazine*, 13(1), 55-78. <https://doi.org/10.1109/MGRS.2024.3439365>.
- [4] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015* (pp. 234-241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28.
- [5] Chen, L. C., Zhu, Y., Papandreou, G., & et al. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 801-818). https://doi.org/10.1007/978-3-030-01234-2_49.
- [6] Zhou, L., Zhang, C., & Wu, M. (2018). D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 182-186). <https://doi.org/10.1109/CVPRW.2018.00034>.
- [7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., & et al. (2022). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- [8] Tao, J., Chen, Z., Sun, Z., & et al. (2023). SEG-Road: A segmentation network for road extraction based on transformer and CNN with connectivity structures. *Remote Sensing*, 15(6), 1602. <https://doi.org/10.3390/rs15061602>.
- [9] Zhu, X., Huang, X., Cao, W., & et al. (2024). Road extraction from remote sensing imagery with spatial attention based on Swin Transformer. *Remote Sensing*, 16(7), 1183. <https://doi.org/10.3390/rs16071183>.
- [10] Li, R., Wang, L., Zhang, C., & et al. (2022). A2-FPN for semantic segmentation of fine-resolution remotely sensed images. *International Journal of Remote Sensing*, 43(3), 1131-1155. <https://doi.org/10.1080/01431161.2022.2033837>.
- [11] Guo, Z., Bian, L., Huang, X., & et al. (2024). DSNet: A novel way to use atrous convolutions in semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35, 3679-3692. <https://doi.org/10.1109/TCSVT.2023.3342127>.
- [12] Ruby, U., & Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(10).
- [13] Sudre, C. H., Li, W., Vercauteren, T., & et al. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 240-248). Springer. https://doi.org/10.1007/978-3-319-67558-9_28.
- [14] Yu, Y., Wang, C. P., Fu, Q., & et al. (2023). A review of semantic segmentation evaluation metrics and evaluation methods. *Computer Engineering and Applications*, 59(6), 57-69.
- [15] Bo, W., Liu, J., Fan, X., & et al. (2022). BASNet: Burned area segmentation network for real-time detection of damage maps in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-13. <https://doi.org/10.1109/TGRS.2021.3106358>.
- [16] Mehta, D., Skliar, A., Ben Yahia, H., & et al. (2022). Simple and efficient architectures for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2628-2636).
- [17] Huang, H., Lin, L., Tong, R., & et al. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1055-1059). IEEE. <https://doi.org/10.1109/ICASSP40776.2020.9053430>.
- [18] Wang, L., Li, R., Zhang, C., & et al. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 196-214. <https://doi.org/10.1016/j.isprsjprs.2022.06.011>.
- [19] Liu, J., & Huang, C. (2024). Multi-scale adaptive road extraction network for high-resolution remote sensing images. *Frontiers in Computing and Intelligent Systems*, 10(3), 101-106.
- [20] Yuan, F., Wang, Z. Y., & Lu, K. (2025). Road extraction from remote sensing images based on DeepLabv3+ with multi-level feature fusion. *Laser Journal*, 46(7), 93-100.