



# OPEN A dynamic attention mechanism for road extraction from high-resolution remote sensing imagery using feature fusion

Haoming Bai, Chao Ren<sup>✉</sup>, Zhenzhong Huang & Yao Gu

Accurate road information is critical for intelligent navigation and urban planning. Compared with traditional road detection methods, deep learning-based approaches have demonstrated significant advantages in road extraction from remote sensing imagery. However, challenges such as occlusion by vegetation and buildings, as well as the similarity between roads and surrounding objects, often lead to incomplete road extraction. To address these issues, we propose a novel deep learning model, RISENet, which consists of three main components: a dual-branch fusion encoder, a multi-layer dynamic spatial channel fusion attention mechanism (MCSA), and a hybrid feature dilation-aware decoder. The dual-branch encoder leverages dual convolutions and multi-head deep convolutions to extract fundamental features and capture fine-grained details. The feature fusion module integrates both global and local information, enhancing the model's ability to represent features effectively. The MCSA captures long-range dependencies within remote sensing images, improving the differentiation between roads and other objects. The dilation-aware decoder dynamically expands the receptive field, preserving global features while reducing the loss of fine details. The proposed RISENet was comprehensively evaluated on three distinct road segmentation benchmarks, demonstrating superior accuracies of 90.04%, 92.24%, and 88.18% respectively. In terms of visual quality and quantitative indicators, the method proposed in this study demonstrates excellent performance. The ablation experiments have also confirmed the effectiveness of the adopted loss function and fusion strategy. These fully indicate that RISENet performs remarkably well in road segmentation tasks across various datasets and exhibits considerable robustness.

**Keywords** Dual-branch fusion encoder, Hybrid feature dilation-aware decoder, Dynamic attention mechanism, Road segmentation, Remote sensing

With the development of modern technology, Earth observation programs have been expanding, leading to a rapid increase in the output of remote sensing images. The quality of high-resolution remote sensing image data has continuously improved, and we have now entered the era of remote sensing big data. Roads are a crucial component of remote sensing imagery. By extracting and analyzing road information, we can make more accurate data-driven decisions in various fields such as autonomous driving<sup>1</sup>, urban management<sup>2</sup>, map updating<sup>3</sup>, traffic navigation<sup>4</sup>, route planning<sup>5</sup>, and agricultural irrigation<sup>6</sup>. This can enhance the efficiency of social operations and resource utilization, thereby promoting the intelligent development of cities and improving quality of life and environmental sustainability<sup>7</sup>.

As both domestic and international experts delve deeper into research on road extraction from high-resolution remote sensing images<sup>8</sup>, Lian et al. proposed that the road features in high-resolution remote sensing image extraction tasks can be classified into several categories: geometric features—roads typically appear as long linear shapes, with little variation in width and curvature within local ranges, and intersections often have geometric forms such as T, Y, or cross shapes; radiometric features—the road edges on both sides exhibit noticeable edge information, and the grayscale distribution of the road surface follows regular patterns, with strong radiometric differences compared to surrounding features; topological features—roads exhibit a distinct network structure, with various types of roads interconnected at intersections; contextual features—these focus on semantic information such as conditions, rules, and evidence, which contribute to prior or posterior

College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541006, China. ✉email: renchao@glut.edu.cn

judgments about roads. This implies that there are certain spatial semantic relationships between roads and surrounding features, which can be used as indirect inference markers for road extraction, such as the strong co-occurrence of roads with trees, buildings, vehicles, road signs, and traffic medians; auxiliary features—roads, as man-made objects, have a wealth of related information that can assist in road interpretation and extraction, such as vector data, DSM, trajectory data, and crowdsourced data<sup>9</sup>.

Based on the aforementioned classification of road features, road extraction methods can be roughly divided into two categories: traditional methods and deep - learning methods. In traditional methods, road feature extraction mainly relies on three techniques: template matching, knowledge - driven approaches, and object - oriented methods. Template matching is the most widely applied in traditional road extraction due to its simplicity and intuitiveness. It detects roads by creating templates with specific shapes and moving them across remote - sensing images to find the best - matching areas. When road features are relatively consistent, this method is effective and is also used by many commercial software to automate road extraction. However, in complex scenarios or when road features vary significantly, it is necessary to adjust template parameters or design multiple templates, which limits its applicability and automation level. Inspired by the breakthroughs of deep learning in fields such as image classification, researchers have applied deep - learning techniques to road extraction. Compared with traditional methods<sup>10</sup>, deep - learning methods have significantly improved the accuracy of road extraction. The deep - learning techniques for road extraction from remote - sensing images mainly include convolutional neural networks (CNNs), Transformer - based models, and specific network structures and their variants, each of which is suitable for different scenarios and tasks.

Convolutional neural networks (CNNs), as one of the core architectures in deep learning, have made significant progress in remote sensing image road extraction tasks. CNN-based methods typically adopt a pixel-by-pixel processing strategy. Wei et al. pointed out that neural networks have strong capabilities when handling large-scale data and can automatically extract abstract semantic features from raw data<sup>11</sup>. While traditional CNNs excel at hierarchical feature extraction and processing large datasets, they tend to lose some spatial information during convolution and pooling operations, especially in deeper layers of the network, which can weaken the model's ability to capture fine details. Furthermore, CNNs have limitations in handling long-range dependencies, as their design primarily focuses on capturing local correlations. Excessive attention to local information can cause the model to converge on a locally optimal solution, which results in poor global perception, leading to discontinuities or misidentification in the extracted roads. Current methods still face challenges in improving road extraction performance in complex terrains and occlusion scenarios<sup>12</sup>. Additionally, the model's reliance on large amounts of labeled data remains a significant challenge, as efficient utilization of limited training resources is crucial<sup>13</sup>. Moreover, the generalization ability of the model needs further enhancement to adapt to different regions and types of remote sensing images<sup>14</sup>. To effectively handle long-range dependencies, more complex network structures or deeper models may be required to achieve this goal<sup>15</sup>. Fortunately, recent studies have shown that the Transformer architecture has great potential in addressing these issues. Its self-attention mechanism can better establish long-range dependencies, facilitating the use of global information from different layers. Therefore, our motivation is to introduce the attention mechanism into the road extraction task, which may further enhance segmentation performance.

This paper proposes a novel RISENet model for accurate road extraction from the complex road structures in remote sensing images. The RISENet model incorporates three key innovations: a dual-branch fusion encoder, a multi-layer dynamic spatial-channel fusion attention (MCSA) bridging module, and a hybrid feature dilation-aware decoder. The dual-branch fusion encoder combines traditional dual-convolution structures for initial feature extraction, while a multi-head deep convolution module captures more detailed and diverse feature information. To enhance the expressive power of feature representations, a feature fusion convolution module is designed to effectively integrate both global and local information, thereby improving the diversity and richness of features. Building on this, the MCSA mechanism introduces a dynamic spatial-channel fusion strategy, which effectively captures long-range dependencies in remote sensing images and dynamically adjusts weights based on the characteristics of different land cover types to optimize the distinction between road and non-road objects. Finally, the hybrid feature dilation-aware decoder expands the receptive field, enhancing the model's attention to global features while minimizing the loss of detailed information. Extensive experimental analysis on the Massachusetts Road dataset, DeepGlobe Road dataset, and LSRV Road dataset shows that RISENet exhibits outstanding generalization, robustness, and stability across these diverse datasets. Through these innovations, we aim to provide new perspectives and solutions for research in the field of road extraction from remote sensing images.

The main contributions of this paper are as follows:

1. This study proposes a novel encoder-decoder architecture named RISENet, which captures long-range dependencies and road detail features to achieve precise road extraction from remote sensing imagery.
2. The dual-branch fusion encoder and hybrid feature dilation-aware decoder are developed. By employing dynamic convolution to parallelly process local details and supplement missing features while acquiring deep semantic information, this framework ensures complete and continuous road boundary delineation.
3. The MCSA mechanism is introduced. By facilitating inter-group interactions among input information, this mechanism enables effective alignment of multi-source contexts, captures cross-modal correlations, and assigns adaptive weights to different regions, thereby guiding and enhancing the RISENet model.

The structure of the remainder of the paper is as follows. Section “[Related works](#)” introduces the related work to this model. Section “[Methods](#)” provides a detailed introduction to the RISENet model architecture and design philosophy. Section “[Results](#)” describes the datasets and experimental details used, followed by ablation

experiments and performance testing across various scenarios, including comparisons with popular models. Section “Conclusions” concludes the paper.

## Related works

In recent years, there have been continuous breakthroughs in deep learning models for road extraction, mainly focusing on improving network architectures and incorporating attention mechanisms to enhance the accuracy and efficiency of road information extraction. To address these challenges, several studies have proposed new solutions. For example, Q. Xu et al. proposed P2CNet, which uses a dual-branch network combined with a gated self-attention module (GSAM) and missing-part loss (MP loss) to significantly improve road extraction accuracy<sup>16</sup>. Z. Gong et al. introduced an improved UNet network, incorporating a pre-trained VGG model, a dense dilated convolution module, and a channel attention module to effectively address occlusion, scale variation, and connectivity issues in remote sensing images<sup>17</sup>. L. Qiu et al. proposed SGNet, which uses a dual-branch backbone network to adaptively fuse dense semantic features with sparse boundary features, significantly enhancing road extraction accuracy<sup>18</sup>. Additionally, R. Lian and L. Huang introduced the DeepWindow method, which combines sliding windows with CNN models to directly track road networks in images, reducing training costs and enhancing the robustness of the method<sup>19</sup>. Z. Yang et al. proposed RCFSNet, which significantly improves road extraction connectivity and accuracy by using a multi-scale context extraction module and full-stage feature fusion module<sup>20</sup>. EfficientNet utilizes a compound scaling model to scale all dimensions of the model, greatly enhancing its efficiency in extracting road information in complex scenes<sup>21</sup>. Similarly, CSPNet reduces redundant gradient information by using a cross-stage local network structure, improving road segmentation accuracy in complex backgrounds while reducing computational load<sup>22</sup>. RADAnet combines a road enhancement module with a deformable attention module to leverage road shape priors and deformable convolution techniques for precise road feature extraction in high-resolution remote sensing images<sup>23</sup>. ECA-Net incorporates an efficient channel attention mechanism to significantly enhance the feature expression ability of deep convolutional neural networks without adding extra computational burden<sup>24</sup>. DPENet enlarges the model's visual range through a multi-scale upsampling mechanism and integrates SE modules into residual blocks, enhancing the semantic reconstruction of road features, demonstrating excellent performance across multiple road datasets. The BEMRF-Net uses the boundary-aware self-attention module to improve boundary detection, and the multi-scale fine fusion module to address the problem of discontinuity<sup>25</sup>. The RMU-Net adopts an extraction method that combines the multi-scale feature and region threshold approaches, which enhances the network's ability to capture scale features and pixel details, thereby improving the accuracy and integrity of extracting target information<sup>26</sup>. The LBA-MCNet balances the boundaries near the objects, solves the problem of inaccurate edge positions, and enhances the collaborative modeling of the background and foreground contexts through affinity learning<sup>27</sup>. Construct a Global-local context-aware (GLCA) module to fully capture the differential features of the target context, and build a SCFA module to enhance the performance of target features at different scales<sup>28</sup>.

Despite significant progress in road extraction methods in terms of accuracy and efficiency, model performance may still decline when handling specific road types or complex remote sensing images, especially in intricate scenarios where accuracy and robustness are compromised<sup>29</sup>. In response to this challenge, researchers have increasingly recognized that integrating various attention mechanisms and innovatively improving these mechanisms can effectively enhance the model's segmentation capability, making it a current research trend.

Traditional spatial-channel attention mechanisms, such as SE, CBAM, and ECA, while capable of enhancing the feature extraction capability of models, still face challenges in terms of computational efficiency, information preservation, and adaptability. The advent of Transformers has revolutionized deep learning models. Since their introduction by Vaswani et al. in 2017<sup>30</sup>, Transformers have become the dominant model architecture in computer vision (CV). Their core innovation, the self-attention mechanism, enables the model to consider all positions in the input sequence simultaneously, effortlessly capturing long-range dependencies without the step-by-step processing required by CNNs<sup>31</sup>. Key components of Transformer models include multi-head attention, stacked layers, positional encoding, residual connections, layer normalization, and encoder-decoder structures. These characteristics collectively empower Transformers to handle complex data<sup>32</sup>. In recent years, various Transformer variants have emerged. For instance, some studies focus on reducing the computational complexity of Transformers using sparse attention mechanisms. Models like Longformer<sup>33</sup> and Sparse Transformer<sup>34</sup> reduce the computational cost of processing long sequences by incorporating sparsity patterns. Additionally, research on Vision Transformers (ViT) has deepened, with subfields such as general ViT<sup>35</sup>, efficient ViT<sup>36</sup>, training Transformers, and convolutional Transformers<sup>37–39</sup>. Significant contributions have been made by W. Lu et al.<sup>40</sup>, who proposed a novel model called RoadFormer, combining the Swin Transformer and spatial-channel separable convolution. This model effectively enhances the capture of long-range information and global context, further advancing the application of Transformers in remote sensing image road extraction. These studies indicate that the Transformer architecture plays a significant role in road extraction tasks from remote sensing images.

Although these achievements have been made, challenges remain in high-precision road extraction tasks. These challenges are primarily caused by the elongated shape of roads, as well as occlusions by trees and buildings, which can result in some roads being invisible in the images. Furthermore, the spectral diversity and similarity of roads increase data complexity, while noise and distortion in road data further complicate the extraction task. Notably, the number of road pixels is much smaller than that of background pixels, leading to an imbalance between positive and negative samples, which challenges the model's learning and generalization ability. Therefore, This paper proposes a novel deep learning model, RISENet which integrates a dual-branch fusion encoder, a multi-head dynamic spatial-channel fusion attention module, and a hybrid feature dilation-aware decoder, aimed at overcoming these challenges.

## Methods

### Overview of method

RISENet’s architecture consists of a five-layer dual-branch fusion encoder and a five-layer hybrid feature dilation-aware decoder. The encoder is composed of several key modules, including the Double Convolution Module (DoubleC), the Multi-Head Depthwise Separable Convolution Module (MDW), and the Feature Fusion Module (FC). The decoder employs a strategy that combines standard convolution operations with feature dilation-aware modules. A Multi-layer Dynamic Spatial Channel Fusion Attention (MCSA) mechanism is added to the connection between the encoder and the decoder. These modules collaborate to enhance the model’s feature extraction and information fusion capabilities, ultimately improving performance.

First, coordinate convolution is introduced into the input remote sensing road images, integrating spatial coordinate information with the input images, thereby enhancing the network’s spatial awareness. The road images with coordinate convolution are processed separately by the MDW and DoubleC modules. At the same level, the feature maps output from both modules are simultaneously passed into the FC module for feature fusion.

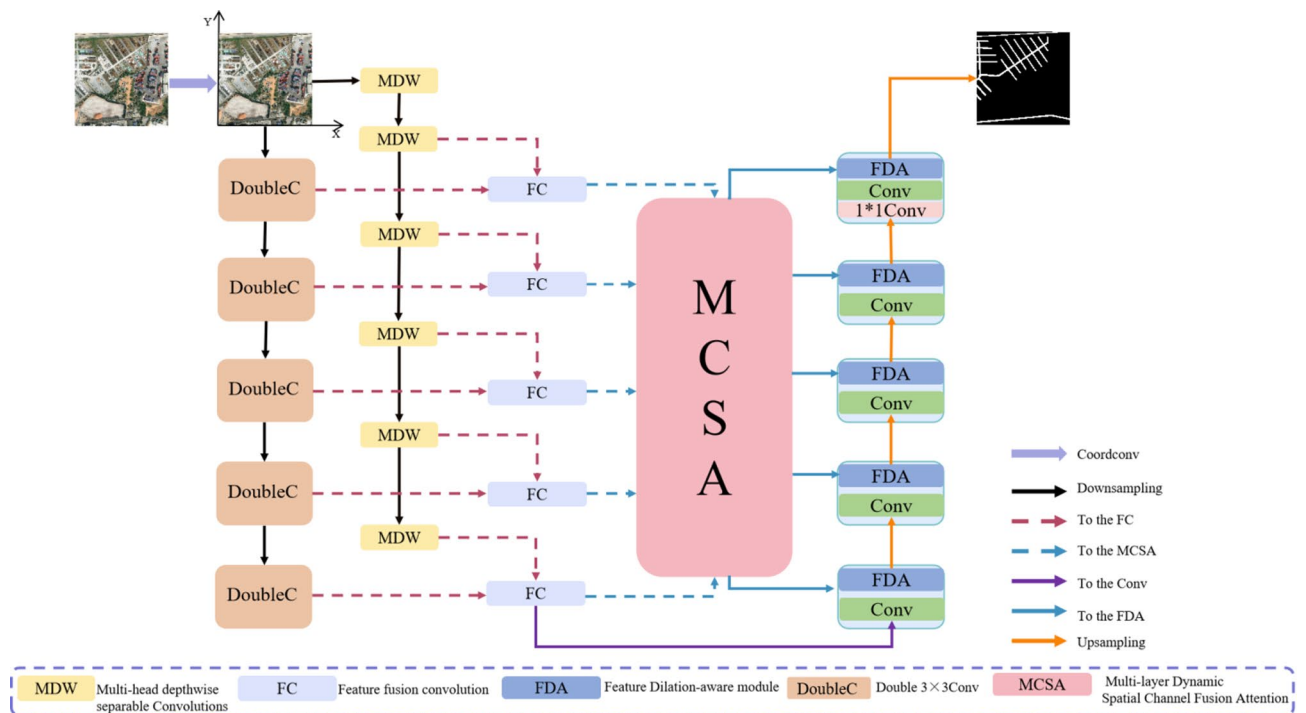
Second, after feature fusion, the output images from each layer are fed into the MCSA mechanism. By maintaining global perception while optimizing local features, the MCSA mechanism allocates different weight information to the background and the road, improving segmentation accuracy. Finally, each layer of the decoder consists of two modules: the Feature Dilation-aware Fusion Module (FDA) and the Convolution Module. The FDA module decodes by gradually expanding the receptive field, utilizing multi-scale road features to obtain more comprehensive contextual information. Figure 1 illustrates the RISENet model’s architecture.

### Coordinate convolution

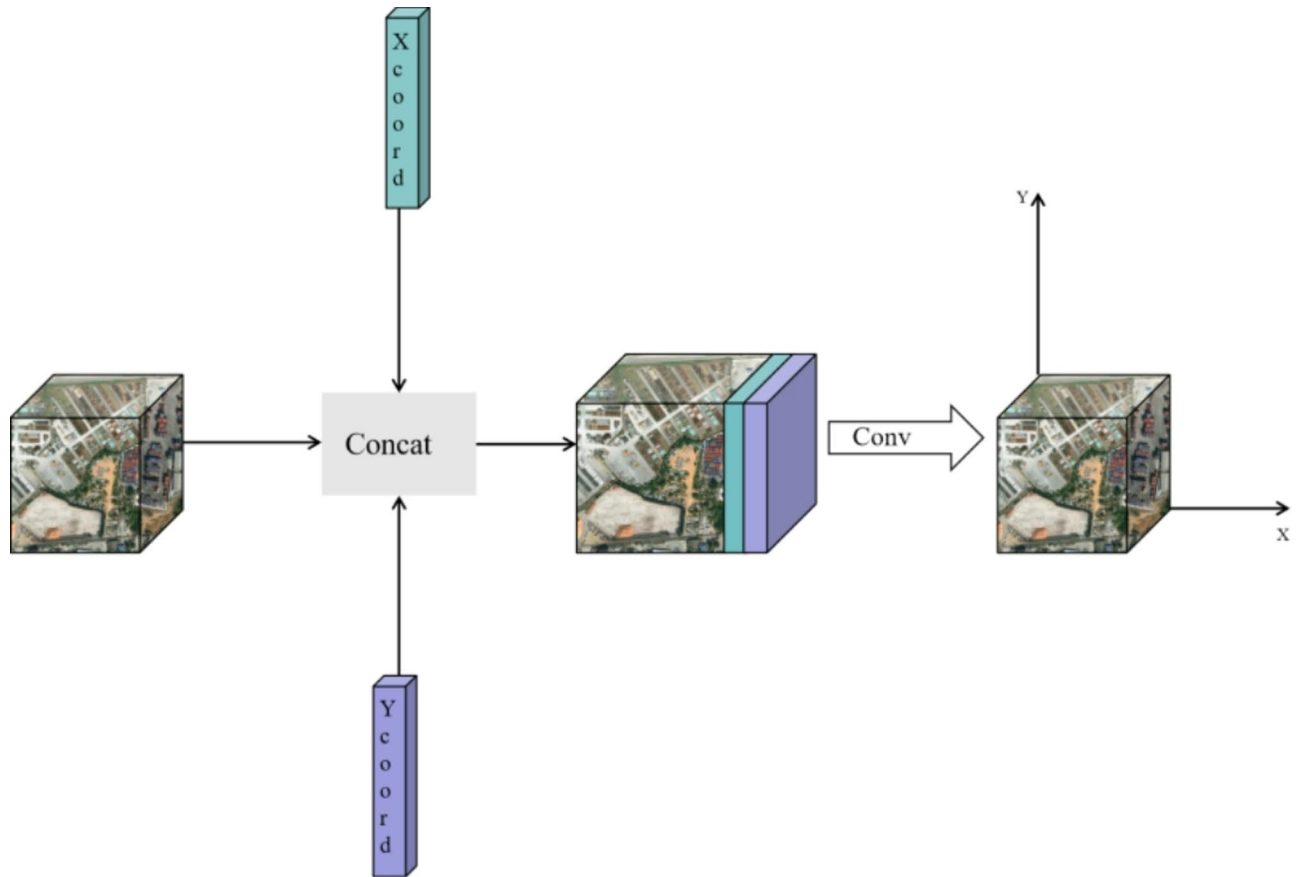
Compared to traditional convolution, where pixel positions are often ignored, CoordConv adds two additional channels to the input feature map: one representing the x-coordinate and the other representing the y-coordinate. This coordinate matrix has the same size as the feature map at the current layer. It not only retains the full translation invariance of traditional convolution but also introduces a degree of translation dependence. Most importantly, it endows the model with spatial awareness capability.

$$Y = \text{Conv}(\text{Concat}(X, \{(X_{ij}, Y_{ij})\})) \tag{1}$$

The horizontal and vertical coordinates in the coordinate matrix  $X_{ij}, Y_{ij}$  represent the coordinates, where  $X$  denotes the input feature vector, “Conv” represents the convolution operation and “Concat” refers to the concatenation of the input feature vector and the coordinate matrix along the feature dimension.  $Y$  represents the output fused with positional information. The structure of CoordConv is shown in Fig. 2.



**Fig. 1.** RISENet overall architecture diagram. Satellite images were generated using the DeepGlobe Road Dataset under CC BY 4.0.



**Fig. 2.** Schematic of the CoordConv structure. Satellite images were generated using the DeepGlobe Road Dataset under CC BY 4.0.

### Multi-head depthwise separable Convolution module

Unlike traditional semantic segmentation, the core objective of remote sensing road segmentation is to identify and segment roads of varying widths. Given the significant scale variation of roads, we introduce a MDW module to effectively address this challenge. The MDW module aims to capture richer contextual information while accurately extracting multi-scale texture features. As a multi-scale processing module with a relatively sparse network architecture, MDW generates dense feature maps. It first utilizes  $3 \times 3$  depthwise separable convolutions to obtain local information, and then employs parallel depthwise separable convolutions with kernel sizes of  $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$  to capture contextual information across multiple scales. The output feature maps from these three convolutions are subsequently summed, reducing feature information loss while fusing the features. Finally, a  $1 \times 1$  convolution is used to integrate local and contextual features. The  $1 \times 1$  convolution acts as a channel fusion mechanism, allowing integration of features with different receptive field sizes. Notably, our MDW module avoids the use of dilated convolutions to prevent overly sparse feature representations. As a result, the MDW module can capture extensive contextual information without compromising the integrity of local texture features. The structure of the MDW module is shown in Fig. 3.

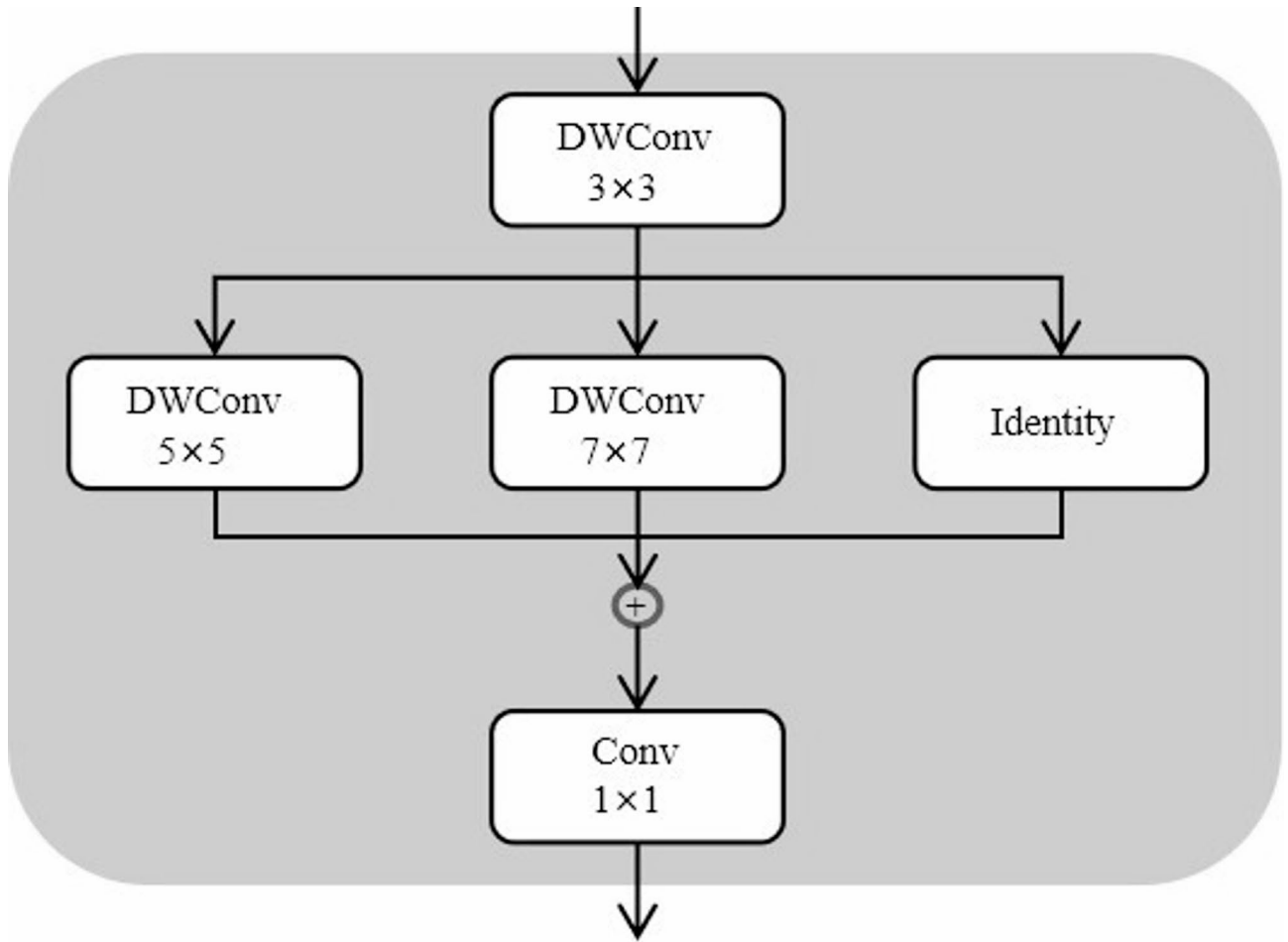
To further clarify the computational process of the MDW module, we first review the mathematical formulation of depthwise separable convolution. Depthwise separable convolution consists of two parts: depthwise convolution and pointwise convolution. For a given input feature map  $X$ , the operation of depthwise separable convolution can be expressed as:

$$Y_{\text{depth}} = \text{depthwise Conv2d}(X, \text{depth\_multiplier} \times K \times K) \quad (2)$$

$Y_{\text{depth}}$  represents the output of the depthwise convolution,  $X$  denotes the input feature map, and  $K$  indicates the size of the convolutional kernel. The operation  $\text{depthwiseConv2d}$  refers to the depthwise convolution, while  $\text{depth\_multiplier}$  is the depth multiplier, which is typically set to 1. The pointwise convolution is computed as shown in Eq. (3):

$$Y_{\text{point}} = \text{pointwiseConv2d}(Y_{\text{depth}}, C_{\text{out}}) \quad (3)$$

$Y_{\text{point}}$  represents the output of the pointwise convolution,  $\text{pointwiseConv2d}$  denotes the pointwise convolution operation, and  $C_{\text{out}}$  is the number of output channels. Therefore, the formula for DWConv can be expressed



**Fig. 3.** Schematic representation of MDW.

asdepthwise Conv2d + pointwiseConv2d.Next, the computation process of the MDW module can be represented as follows:

$$L_{S+1} = Conv_{K \times K}(X_S)$$

$$Z_{S+1}^{(m)} = DWConv_{K^{(n)} \times K^{(n)}}(L_S), m = 1, 2, 3, n = 3, 5, 7 \tag{4}$$

Here,  $L_{S+1}$  represents the local features extracted after convolution,  $X_S$  is the feature map at the input of layer  $S$ , and  $Z_{S+1}^{(m)}$  denotes the multi-scale contextual features extracted by the  $m$ -th convolution kernel through depthwise separable convolution for  $K^{(n)}$ . In the MDW module,  $n = 2(m+1)+1$ , which indicates that  $n$  is proportional to  $m$ . As  $m$  increases,  $n$  also increases, and  $K^{(n)}$  varies with  $m$ . The combination of different convolution kernels can capture contextual features at different scales as  $m$  changes.

$$F_S = Conv_{1 \times 1} \left( L_S + \sum_{m=1}^3 Z_S^{(m)} \right) \tag{5}$$

This represents the feature map output by the MDW at layer  $S$ . On a macro level, the generated feature maps  $F_1$ ,  $F_2$ , and  $F_3$  are the outputs after applying  $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$  convolution kernels, respectively. The final output is obtained by summing and fusing with a  $1 \times 1$  convolution, and is represented as:

$$F_f = (F_1 + F_2 + F_3) \times Conv_{1 \times 1} \tag{6}$$

$F_f$  is the final output feature map of the MDW module. This multi-scale depthwise separable convolution design effectively captures contextual information at different scales, thereby enhancing the model's expressiveness and robustness.

### Multi-layer dynamic spatial channel fusion attention mechanism

This chapter introduces an innovative attention mechanism—Multi-layer Dynamic Spatial-Channel Fusion Attention (MCSA), specifically designed for road extraction tasks in remote sensing images. MCSA adaptively learns key information from skip connections, reduces the influence of irrelevant information, and improves the segmentation accuracy of fine-grained details. The module consists of three core components: Multi-Head Embedded Patches (MEP), Multi-layer Dynamic Spatial Attention (MDSA), and Multi-layer Dynamic Channel Attention (MDCA).

MCSA divides the input features into several small patches using the MEP module, followed by linear mapping. The patches are then processed separately by the MDSA and MDCA modules to compute the query (Q), key (K), and value (V) vectors of the features. Both modules process the features in different ways, but the final fusion is carried out through weighted sums, convolutions, and residual connections to ensure that features at different scales are fully utilized while maintaining a balance between global and local information. MDCA mainly operates on the channel dimension, normalizing the input features first, then directly mapping to queries (Q), while keys (K) and values (V) are generated through a linear transformation after feature concatenation. Subsequently, Q and K are processed with matrix operations and normalized by Softmax to enhance the interaction of information between channels. MDCA uses residual connections to ensure that high-frequency information is highlighted while maintaining the integrity of low-frequency features and enhancing the model's ability to capture long-range dependencies. MDSA mainly works on the spatial dimension, and the order of calculating Q and K differs from MDCA—normalization is performed before concatenation, followed by linear mapping, making it more suitable for capturing both local and global spatial features. As MDSA is positioned at the backend of the MCSA structure, its output is influenced not only by the weights computed within the module but also by the channel information extracted by MDCA. Specifically, the attention weights computed by MDSA are applied to the input features, and these features are fused with the outputs from MDCA through convolution, with further enhancement via residual connections. The combination of the two modules ensures that MCSA retains multi-scale channel information while optimizing spatial expression capability, ultimately improving the model's ability to capture complex features and excelling in remote sensing image road analysis tasks. The MCSA module is illustrated in Fig. 4.

#### Multi-head embedded patch module

The Multi-head embedded Patch module (MEP) is composed of two modules, PAG and DWP. The PAG module processes 2D images, performing average pooling downsampling on image  $x \in \mathbb{R}^{H \times W \times C}$  and reshaping the image into a flattened sequence of patches, denoted as  $\mathbb{R}^{N \times (P^2 \cdot C)}$ , where H is the resolution of the original image, C is the number of channels in the original image (for RGB images,  $C=3$ ), P is the resolution of each image patch, and  $N = HW/P^2$  represents the number of image patches produced. Subsequently, N small patches are fed into the DWP module to generate T tokens, where  $t_i$  denotes each feature vector.

$$T = \text{DWP}(\text{Reshape}(\text{AvgPool2D}(X))) \quad (7)$$

$$T = \{t_1, t_2, t_3, t_4, \dots, t_n\} t_i \in \mathbb{R}^d \quad (8)$$

The primary function of depthwise separable convolution in the MEP module is to divide the input image into fixed-size patches, where each patch represents a local region of the image. After undergoing depthwise separable convolution, these patches extract features and are transformed into embedding vectors. In road extraction tasks, the shape, direction, and surrounding environment of roads may exhibit significant variations. This processing method is crucial for MCSA, as it enables the model to comprehensively consider both global and local information regarding the road morphology in remote sensing imagery, thus enhancing its road recognition capabilities. The image processing schematic is illustrated in Fig. 5.

#### Multilayer dynamic channel attention

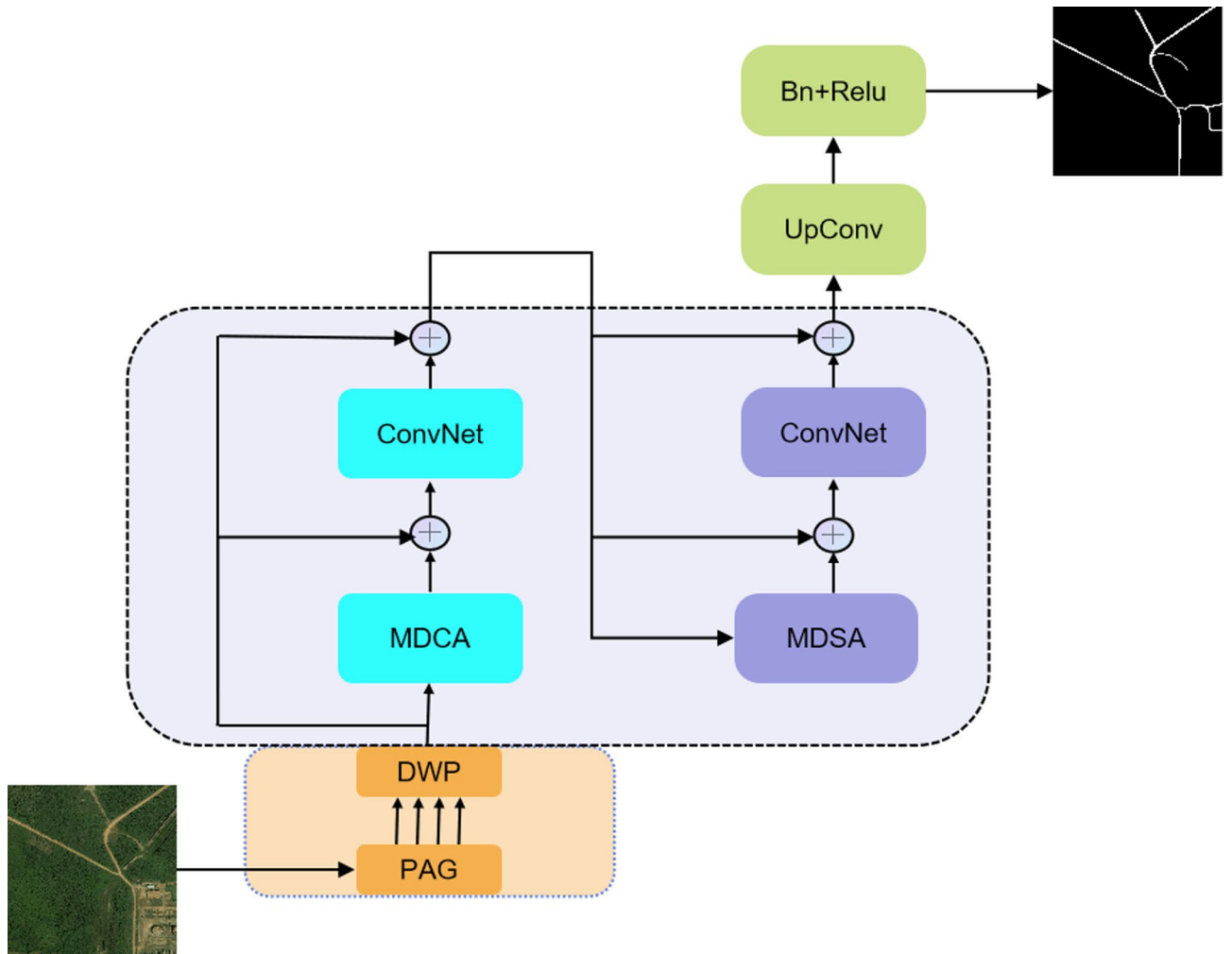
To address the limitations of traditional spatial attention mechanisms, such as SENet, when processing multi-scale information, we propose the Multilayer Dynamic Channel Attention (MDCA) module. This module is primarily used to capture features at different scales, enhancing the model's focus on both local details and global contextual information. By utilizing a residual structure, it prevents the loss of feature information, highlights high-frequency components, and facilitates the interaction and fusion with low-frequency features. Additionally, it effectively captures long-range dependencies in the data, thereby improving the model's understanding of road information. After the remote sensing image undergoes MEP processing, a sequence consisting of multiple tokens is obtained. This sequence is then fed into our MDCA, where each token is processed through LayerNorm and results in  $\hat{t}_i = \text{LayerNorm}(t_i)$ . Within the MDCA, the features are first normalized and then directly mapped to generate the Query (Q) values. The formula for generating Q is as follows.

$$Q_i = W_Q \hat{t}_i + b_Q \quad (9)$$

For the generation of the Key (K) and Value (V), after normalization, the tokens are first concatenated and then undergo a linear transformation to generate the final K and V. The formulas for generating K and V are as follows.

$$\hat{T} = [\hat{t}_1; \hat{t}_2; \dots; \hat{t}_n] \quad (10)$$

$$K_i = W_K \hat{T} + b_K \quad (11)$$



**Fig. 4.** Schematic of the MCSA structure, the bottom four small arrows represent the corresponding output of the encoder. Satellite images were generated using the DeepGlobe Road Dataset under CC BY 4.0.

$$V_i = W_V \hat{T} + b_V \tag{12}$$

Here,  $\hat{T}$  refers to the vector after concatenation.

$W_Q W_K W_V \in \mathbb{R}^{d \times d_k}$  represents the weight of the linear transformation.  $b_Q, b_K, b_V \in \mathbb{R}^{d_k}$  is the bias term. To leverage the attention along the channel dimension, we use the transpose of Q and K. This leads to the core formula of MDCA, as shown below:

$$\widehat{MDCA}(Q, KV) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) \times V \tag{13}$$

Here,  $d_k$  refers to the dimension of the key, and dividing by  $\sqrt{d_k}$  is used to stabilize the gradients. The final generated  $\widehat{MDCA}$  is added to the result of the original LayerNorm through a residual connection.

$$MDCA = \widehat{MDCA} + \hat{t}_i \tag{14}$$

The final output is as follows:

$$Y = DWP(MDCA) \tag{15}$$

In this way, MDCA is able to capture the relationships between different elements in the sequence, and through the Softmax function, it ensures that the output attention weights are normalized, meaning that the sum of all weights equals 1. These weights are then added to the tensor initially output by the LayerNorm operation to improve numerical accuracy. The resulting weighted feature map is then used as input to the ConvNet operation, and the output is subsequently added to the feature map that was originally input into the MDCA module, forming a residual structure. This approach helps the model select more relevant features from the large set of



**Fig. 5.** Input a 512\*512 remote sensing road image, after PAG module, the image is averagely cut into N small images, and then these images are transformed into T embedded vectors through DWP module, and finally these embedded vectors are passed into MCSA. Satellite images were generated using the DeepGlobe Road Dataset under CC BY 4.0.

features while ignoring irrelevant ones, thereby enhancing the model’s generalization ability. The structure of the MDCA is shown Fig. 6a.

*Multi-layer dynamic spatial attention*

Unlike the MDCA model, the MDSA directly maps the normalized features to generate the value (V); for the query (Q) and key (K), however, the normalization is performed first, followed by concatenation of the two normalized features, and then the mapping operation. Since MDSA is placed at the end of the attention module, we have added a KAN network after the DWP module to enhance the model’s ability to capture complex functional relationships. This processing approach allows the MDSA model to handle features more precisely and efficiently. The formula is as follows.

$$Q_i = W_Q \hat{t}_i + b_Q \tag{16}$$

$$K_i = W_K \hat{T} + b_K \tag{17}$$

$$K_i = W_K \hat{T} + b_K \tag{18}$$

$$V_i = W_V \hat{T} + b_V \tag{19}$$

Here,  $\hat{T}$  refers to the vector after concatenation.  $W_Q W_K W_V \in \mathbb{R}^{d \times d_k}$  represents the weight of the linear transformation.  $b_Q, b_K, b_V \in \mathbb{R}^{d_k}$  is the bias term. To leverage the attention along the channel dimension, we use the transpose of Q and K. This leads to the core formula of MDSA, as shown below:

$$MDSA(\widehat{Q}, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V \tag{20}$$

Here,  $d_k$  refers to the dimension of the key. The result obtained by adding the output of  $MDSA(\widehat{Q}, K, V)$  after the residual connection and the original LayerNorm processing.

$$MDSA = \widehat{MDSA} + \hat{t}_i \tag{21}$$

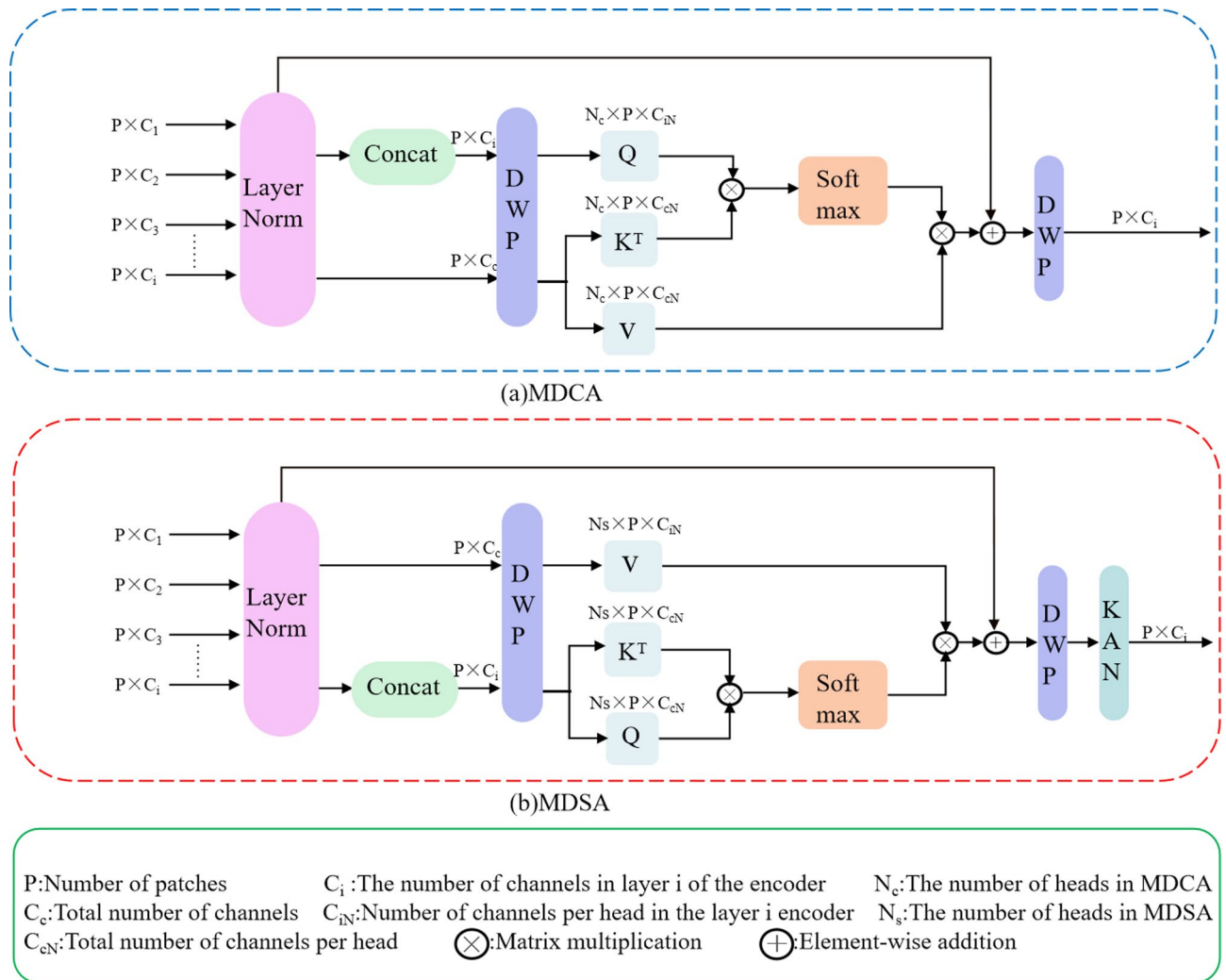


Fig. 6. Schematic diagram of the MDCA structure.

Therefore, the final output is:

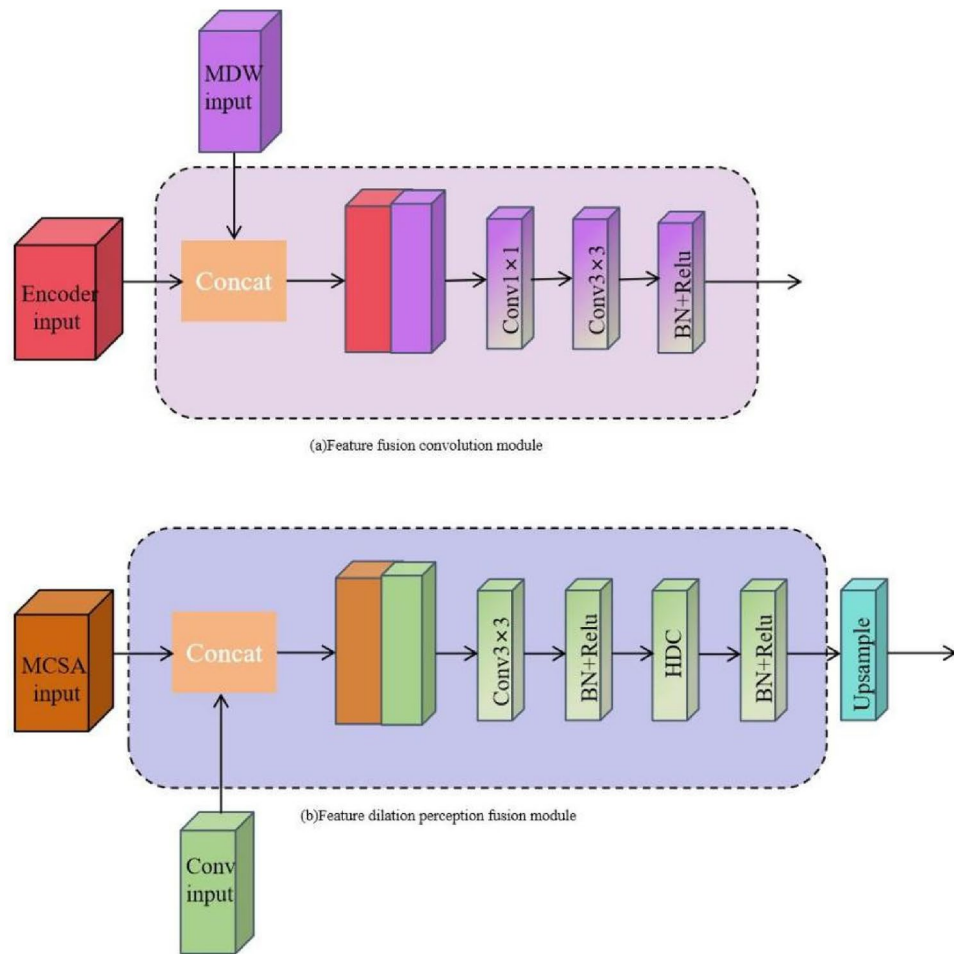
$$Z = \text{KAN}(\text{DWP}(\text{MDSA})) \tag{22}$$

The feature map, after the weighted values are mapped, undergoes convolution and is then added to the initial output feature map of the MDCA. This operation facilitates the fusion of multiple features, allowing for the acquisition of more detailed information. The structure of MDSA is shown in Fig. 6b.

### Feature fusion Convolution module and feature dilatation-aware fusion module

In this chapter, we design two modules: the Feature-Fusion Convolution Module (FC) and the Feature Dilatation-Aware Fusion Module (FDA). FC is placed before MCSA and first receives five layers of images output from DoubleC as Input 1, and the output of the MDW module as Input 2. Then, Input 1 and Input 2 are fused through a concatenation operation and feature extraction is performed using a  $3 \times 3$  convolution followed by a ReLU activation function. The output of this module will serve as the input to MCSA for further allocation of attention weights. The specific structure is shown in Fig. 7a.

Remote sensing images contain a large amount of spatial information and features, and extracting valuable features from them has always been a challenge. To improve image quality and detail, we generate higher-resolution images through convolution operations, which integrate the detailed textures from fine images with the ground change information from coarse images. Effectively extracting these features is crucial for generating high-resolution images. However, the limitation of traditional convolution operations lies in their constrained receptive field and single-layer convolution scale, which typically extract features at a single scale. This limits the ability to capture multi-scale information, potentially leading to the omission of key details and the inability to obtain global context. Therefore, we have designed the Feature Dilatation-Aware Fusion Module (FDA) by introducing Hybrid Dilated Convolution (HDC) technology to expand the receptive field. This allows for



**Fig. 7.** Structure diagrams of FC and FDA.

efficient processing of multi-scale features, ensuring that valuable information in the image is fully captured and preventing information loss, thereby extracting more complete and detailed features.

In the task of road extraction from remote sensing images, traditional road extraction methods have limited capabilities in processing multi-scale features. When faced with complex situations such as variations in road width and similarities between road features and those of the surrounding environment, misjudgments and omissions are likely to occur. However, the Feature Dilatation - Aware Fusion Module (FDA) can effectively address these issues. Its inputs include four layers of images output from the Multi - layer Dynamic Spatial - Channel Fusion Attention (MCSA) (Input 1) and the feature map processed by the previous - layer decoder (Input 2). These two groups of inputs are fused through a concatenation operation and then undergo feature extraction via a series of Hybrid Dilated Convolution (HDC) modules combined with ReLU activation functions. During the feature extraction process, the FDA module avoids weight sharing, enabling each feature image to extract more representative features based on its specific characteristics, providing stronger support for subsequent processing. After completing the feature extraction, the FDA module uses bilinear interpolation for upsampling to enlarge the image size. This processing effectively combines features at different levels and integrates various types of information. On the one hand, by concatenating the features output from the MCSA and those of the previous - layer decoder, it integrates semantic and spatial information at different levels. This allows the model to obtain more comprehensive contextual information when dealing with areas where roads have similar textures to their surroundings, avoiding misjudging non - road areas as roads. When processing roads of different scales, the fusion of multi - scale features can better adapt to changes in road width and improve the recognition accuracy of road boundaries. On the other hand, compared with some simple upsampling methods, the bilinear interpolation adopted by the FDA module can enlarge the image size more smoothly, reducing information loss and image blurring caused by upsampling, and further improving the accuracy of road extraction. Finally, the output of this module serves as the input to the next convolution module for further feature extraction.

In the design of HDC, we not only enhance feature representation and address potential gradient issues in high-level convolutions, but also significantly expand the network's receptive field, improving denoising performance. Compared to traditional dilated convolutions, which typically use a single dilation rate and can result in discontinuous pixel connections in each layer, leading to the “grid effect,” our design employs non-continuous dilation rates. This ensures that the receptive field's coverage area is continuous and more complete.

This innovation greatly improves the model's adaptability when processing road features at different scales, enhances feature fusion, and strengthens the model's generalization ability. The overall structure is shown in Figs. 7b and 8.

## Results

### Dataset

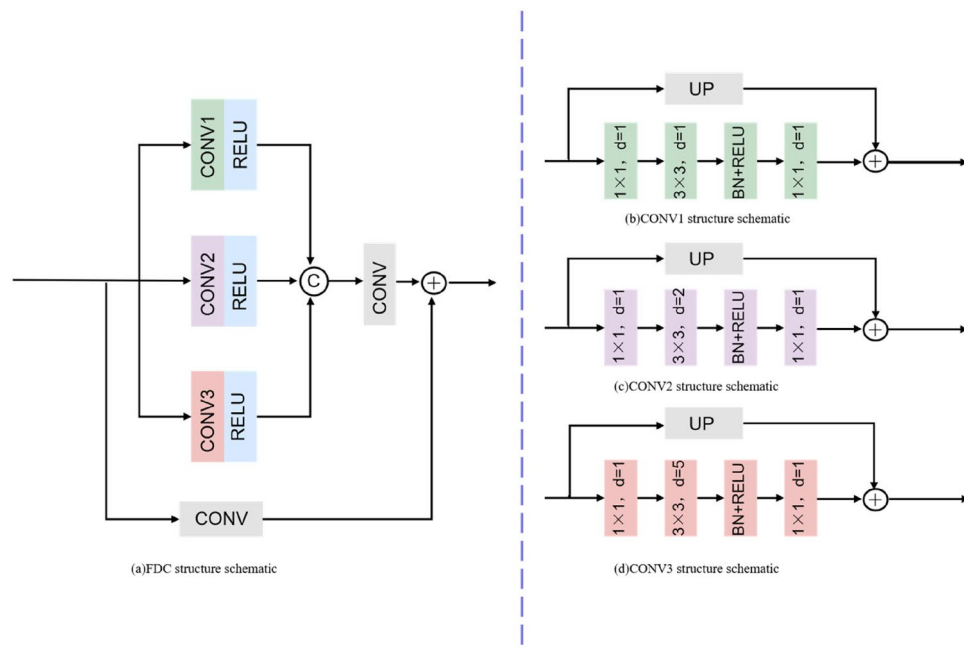
To validate the performance of the RISENet model in terms of both accuracy and generality, we tested it on three different road datasets: the Massachusetts road dataset contains various road types, such as highways, rural dirt roads, and asphalt roads, as well as other elements resembling road features, such as rivers and railways. It includes 1171 aerial images, each with a resolution of  $1500 \times 1500$  pixels, covering a geographic area of approximately 2.25 square kilometers. These images are divided into 1108 images for training and 14 images for validation. The dataset covers various terrains, including urban, suburban, and rural areas, with a total area of over 2600 square kilometers. To enhance the scientific value, we further subdivided the dataset into training and validation sets. The training set consists of 1108 images and their corresponding label images, which were cut into 4432 smaller images of  $512 \times 512$  pixels. These smaller images were randomly allocated in an 8:2 ratio for training and validation, expanding the training set to 3545 images of  $512 \times 512$  pixels, while the validation set contains 887 images of the same size.

The DeepGlobe road dataset provides high-resolution images from Thailand, Indonesia, and India, with a ground resolution of 0.5 m per pixel and an image size of  $1024 \times 1024$  pixels. The dataset is divided into 6226 training images, 1243 validation images, and 1101 test images. To enhance the richness and diversity of the training dataset, we generated image patches of  $512 \times 512$  pixels and randomly assigned them to the training and validation sets in an 8:2 ratio. The annotations for both the training and validation sets are in a binary classification format, where road pixels are labeled as 1, and non-road background pixels are labeled as 0.

The LSRV road dataset includes images from various sources, including aerial images of Boston and its surrounding cities in the United States, Birmingham in the United Kingdom, and Shanghai in China, making the dataset diverse in terms of geographic area and image resolution, allowing for comprehensive testing of the model's generalization ability. The LSRV dataset contains three large-scale images, all collected from Google Earth and precisely annotated for model evaluation. Notably, the distribution of road objects in Shanghai differs significantly from that in Boston and Birmingham. In Shanghai, buildings are taller and denser, and there are many narrow roads between buildings, presenting additional challenges for road detection models. For dataset splitting, we adopted an 8:2 ratio, dividing the dataset into training and validation sets. This strategy ensures sufficient samples for training while allowing for effective model evaluation and optimization through the validation set.

### Training environment

During the experimental development process, Python was selected as the primary programming language, with JetBrains PyCharm 2024 used as the integrated development environment (IDE) for Python. The deep learning framework employed was PyTorch 1.8, running on a system equipped with an Nvidia GeForce RTX 3050 GPU and an Intel® Core™ i5 10th generation processor. The following model parameters were set during training: a



**Fig. 8.** FDC structure schematic.

batch size of 4, 100 training epochs, and an initial learning rate of 0.0001. The Adam optimizer was used to optimize the network. Every 10 epochs, the learning rate was decayed by a factor of 0.1, and a model evaluation was performed.

### Evaluation metrics

To better assess the overall performance of our RISENet in the road extraction task, we employed a binary classification approach. The remote sensing image was classified into two categories: background and road. When the actual class is road, a correct identification as road is considered a True Positive (TP), while an incorrect identification as background is a False Negative (FN). Similarly, when the actual class is background, identifying it as road is a False Positive (FP), and identifying it correctly as background is a True Negative (TN). A schematic of the sample categories is shown in Fig. 9.

To comprehensively evaluate the performance of the model, three key metrics were employed in this study: Precision (P), Recall (R), Intersection over Union (IoU), and F1-score. These metrics provide insights into the model's accuracy, completeness, and overall performance in road extraction from different perspectives.

Precision (P) reflects the reliability of the model's predictions of positive samples. It represents the proportion of truly positive samples among those predicted as positive. A high precision means that the model rarely misclassifies negative samples as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (23)$$

Recall (R), measures the proportion of actual positive samples correctly identified by the model. It focuses on the model's ability to capture all positive samples. A high recall indicates that the model can detect most positive samples, possibly at the cost of precision.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (24)$$

Intersection over Union (IoU): IoU measures the overlap degree between the predicted and ground - truth regions. Widely used in object detection and image segmentation tasks, it calculates the ratio of the intersection to the union of the predicted positive region and the actual positive region.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (25)$$

F1-Score: The F1-score is the harmonic mean of Precision and Recall. It is particularly useful for datasets with imbalanced class distributions, where the number of positive and negative samples is not equal, as this imbalance can affect the model's performance evaluation. By combining both Precision and Recall, the F1-score provides a balanced perspective for assessing the overall effectiveness of the model.

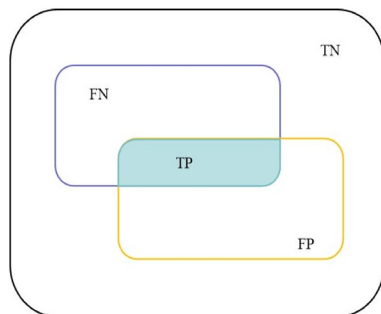
$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (26)$$

### Comparative experiments and results analysis

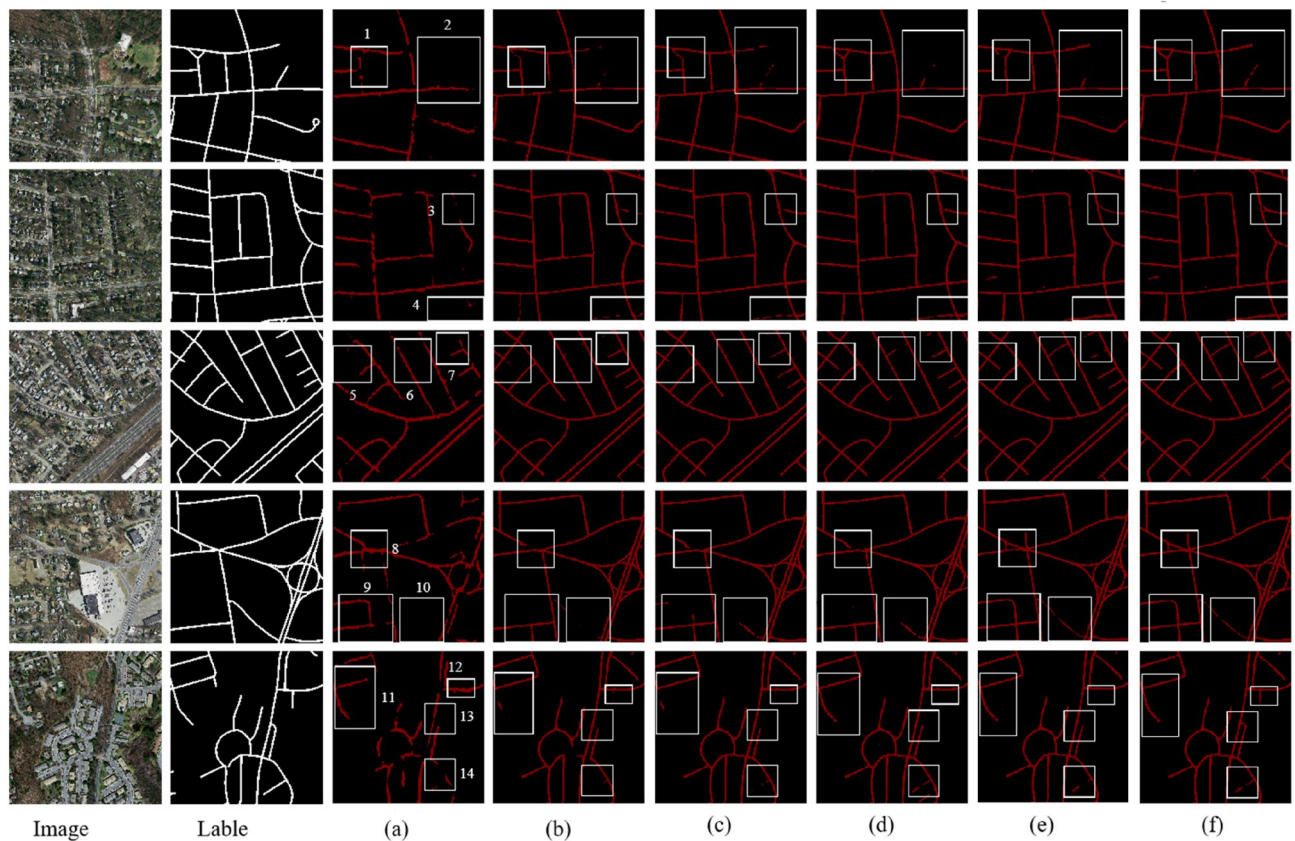
#### *Experimental results on the Massachusetts road dataset*

In this section, we conduct an in-depth analysis of the performance of several algorithms on the Massachusetts Road Dataset, with a particular focus on the characteristics of different regions. To this end, we carefully selected five images as samples, which cover a range of challenges, including unclear local roads and road color/textures that are similar to other features in the environment. We compare our model with UNet<sup>41</sup>, KanUNet, DenseUNet<sup>42</sup>, MDCGANet<sup>43</sup>, and SCSM<sup>44</sup> through a series of comparative experiments. The qualitative experimental results are shown in Fig. 10.

When road segments in remote sensing images are unclear, it poses significant challenges to road extraction. Unclear roads not only affect the recognition of road features but can also lead to erroneous interruptions or



**Fig. 9.** Diagram of the categories of positive and negative samples.



**Fig. 10.** Experimental results on the Massachusetts road dataset. We selected 14 regions for display using white boxes.

discontinuities in the road extraction process. UNet (a) struggles with connecting the roads in unclear areas, resulting in multiple discontinuities in the extracted roads. KanUNet (b), with its compact contextual framework and the addition of the unique Kan module to the UNet model, better addresses some of the unclear road issues. DenseUNet (c), where each layer is directly connected to all previous layers, promotes feature reuse, but still remains susceptible to the influence of unclear road sections. MDCGANet (d), with its multi-scale contextual awareness module, can effectively analyze and classify pixel blocks, resulting in relatively complete road predictions. The SCSM (e) model reconstructs the ordinary attention mechanism by adopting the scene coupling and local - global semantic mask strategies, which can effectively reduce the omission and mis-extraction of road information. RISENet (f), through the combination of multi-layer spatial channel residual attention mechanism and feature expansion perception module, significantly enhances the ability to address this problem, offering a more comprehensive extraction.

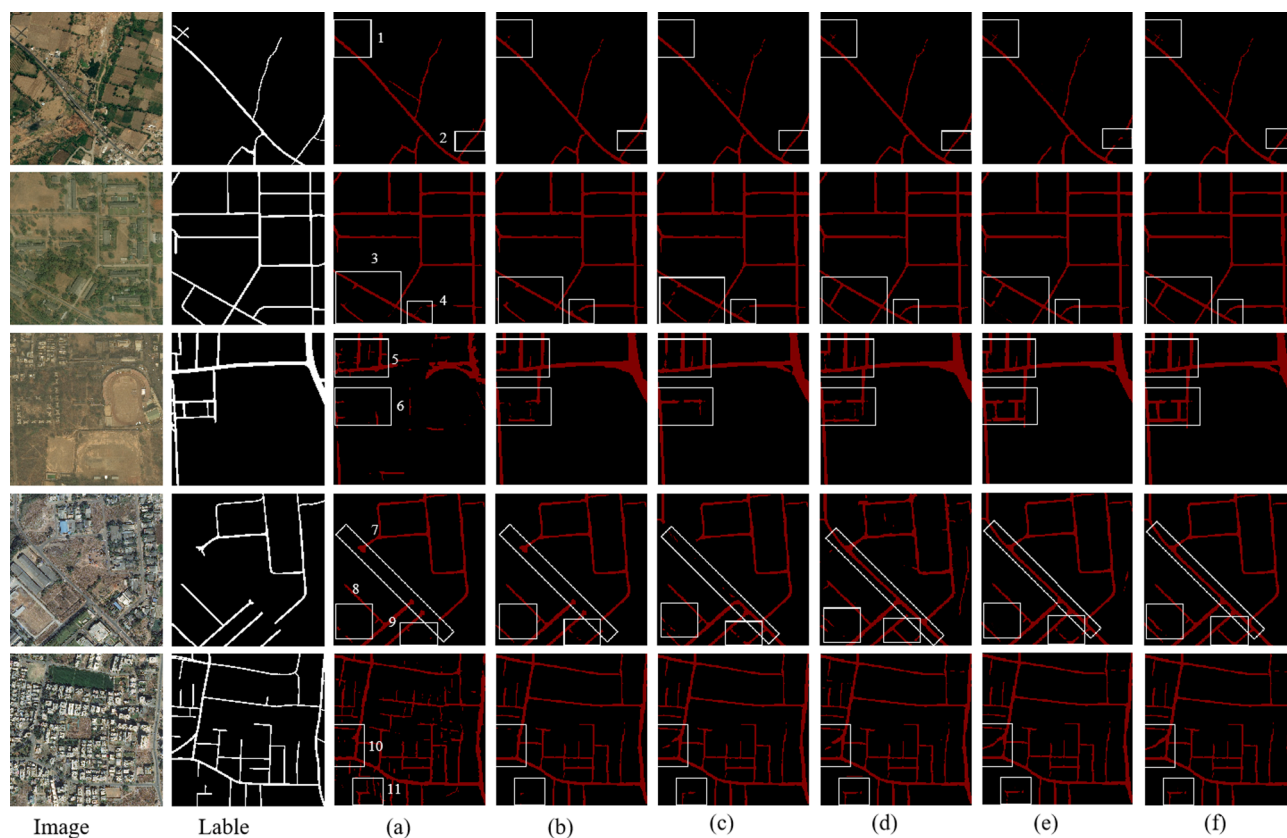
When road color and texture are similar to other features, UNet struggles in recognition without an attention mechanism, leading to many incorrect extractions. KanUNet (b) benefits from its residual connections and the unique Kan network, which helps differentiate roads from similar features, though errors still occur in some regions. DenseUNet (c), despite some enhancement, is still affected by overly similar features. MDCGANet (d), with its local attention mechanism, performs well in distinguishing roads from similar features. When facing extremely complex scenarios, the SCSM (e)'s performance may still be affected to some extent. In this dataset, RISENet (f) stands out with its unique feature perception module and the integration of global and local attention mechanisms, demonstrating more outstanding performance than other benchmark models, especially showing significant advantages in the precise recognition of urban roads.

In the quantitative analysis phase, detailed evaluation results are presented in Table 1.

In the comprehensive evaluation across four key metrics, our proposed RISENet has demonstrated significant superiority in all aspects. MDCGANet is enhanced by directional information and global attention flow, so its accuracy is higher than that of KanUNet which only adds the KAN network. The accuracy of RISENet reached 90.04%, which is 1.36% points higher than that of MDCGANet at 89.57%. Since RISENet has an attention mechanism that can capture long-range dependencies, the F1-score of RISENet is 87.98%, which is far higher than the 80.86% of KanUNet and 1.7% higher than that of SCSM. This demonstrates its outstanding performance in balancing precision and recall. Furthermore, RISENet achieved the highest recall rate at 86.03%, outperforming all other models. This indicates that RISENet has the highest ability to identify all road pixels, with minimal omissions. Although MDCGANet's IoU (Intersection over Union) of 77.15% is slightly lower than RISENet's 82.01%. RISENet still maintains a high level of performance. Overall, RISENet exhibits significant

Scheme	Network	Precision (%)	Recall (%)	IoU (%)	F1-Score (%)
(a)	UNet	86.75	78.36	70.42	82.30
(b)	KanUNet	87.38	75.33	77.02	80.86
(c)	DenseUNet	88.21	78.68	<b>83.01</b>	83.20
(d)	MDCGANet	89.57	82.87	77.15	86.12
(e)	SCSM	89.66	84.19	80.38	86.28
(f)	RISENet	<b>90.04</b>	<b>86.03</b>	82.01	<b>87.98</b>

**Table 1.** Quantitative analysis results of the five models on the Massachusetts road dataset. Significant values are in [bold].



**Fig. 11.** Experimental results on the DeepGlobe road dataset. We selected 11 regions for display using white boxes.

advantages in both qualitative and quantitative evaluations. These experimental results fully demonstrate that RISENet not only excels in efficiency but also boasts high accuracy and recall in road extraction tasks.

#### Experiment results on the deepglobe road dataset

In certain phases of this study, we conducted an in-depth performance analysis of various algorithms on the DeepGlobe Road Dataset to evaluate their capability in road extraction tasks. We specifically selected three representative regional environments, each introducing unique challenges. Figure 11 presents the segmentation results for these regions.

Image 1, 2, and 3 represent the desert plain area, where roads often exhibit narrow characteristics. Some road sections may be covered by sand, causing the road color to blend with the surrounding environment, making it difficult to distinguish. Both DenseUNet (c) and MDCGANet (d) exhibit errors in road recognition, such as discontinuous recognition and failure to correctly identify roads. In particular, UNet (a) failed to extract many roads in Image 3, and the roads exhibit significant discontinuities.

Image 4 represents the industrial suburb area. Roads in industrial suburbs typically have the following characteristics: they appear as elongated, continuous gray or black regions, with a noticeable difference in color intensity compared to the surrounding ground cover. The recognition results of MDCGANet (d) and SCSM (e) are similar, with only partial roads being identified, and SCSM (e) fails to correctly identify all the roads.

RISENet (f), which owns a dual-branch fusion encoder and a hybrid feature dilation-aware decoder, can analyze remote sensing images according to the captured details and global information, precisely distinguishing roads from other ground objects. And it has shown remarkable advantages in this area. For instance, in the manually annotated reference image, there is a narrow and relatively wide road that was not labeled, but our model, after training, accurately identified this road while other models failed to do so.

Image 5 represents the urban center area, where tall buildings are densely distributed, and roads are easily obstructed by buildings and their shadows. Some smaller roads within the buildings may not be effectively identified. KanUNet (b) and UNet (a) did not recognize the curved road in the middle-left building complex of the image, while DenseUNet (c), MDCGANet (d), and SCSM(e) managed to capture this road, although not completely and with some fragmented road sections. Overall, our proposed RISENet (f) has demonstrated considerable road extraction ability under diverse terrain conditions and challenging scenarios. Although there is still potential for further optimization in certain details, RISENet (f) has performed at a level that is comparable to or exceeds existing network models in most test cases. This indicates its considerable application value and development prospects in handling complex road environments.

In the quantitative analysis of the DeepGlobe road dataset, the comprehensive evaluation of four key metrics effectively reflects the robustness of the model. The specific results are shown in Table 2 below.

As shown in the table above, UNet, as a classic architecture in convolutional neural networks, demonstrates robust performance in semantic segmentation, despite not achieving the best results on this dataset. Its precision, intersection-over-union (IoU), recall, and F1 score are 88.81%, 77.80%, 75.26%, and 81.64%, respectively. However, compared to models that incorporate new mechanisms, such as DenseUNet and MDCGANet, UNet's performance lags behind. When compared with UNet, DenseUNet and MDCGANet demonstrate superior performance, especially in terms of the F1-score. UNet has a relatively low F1-score, while MDCGANet and DenseUNet perform more impressively. Specifically, MDCGANet's F1-score reaches 87.12%, and DenseUNet's is 85.32%, with MDCGANet having a slight edge. When further comparing SCSM and MDCGANet, it can be found that the performance of the two models shows different characteristics. Since the scene information of SCSM is decomposed by the scene coupling module and embedded into the attention affinity process to effectively utilize the intrinsic spatial correlation of features, the accuracy of SCSM is 0.35% higher than that of MDCGANet, but in terms of the F1-score, SCSM is 0.61% lower than MDCGANet, this improvement can likely be attributed to the multi-scale feature fusion and conditional generative adversarial network (CGAN) structure utilized by MDCGANet, which helps capture finer details and contextual information in the image. The RISENet proposed in this study outperforms all other models across all metrics, achieving a precision of 92.24%, a recall of 85.72%, an IoU of 82.77%, and an F1-score of 88.86%. These results highlight the exceptional robustness of RISENet in detecting roads in complex environments and refining their boundaries. Overall, RISENet shows significant performance improvement over other networks on the DeepGlobe road dataset, emphasizing the importance of adopting advanced network architectures for road extraction tasks. The high performance of RISENet not only demonstrates its effectiveness in road extraction but also opens possibilities for future applications in more complex scenarios.

#### Experiment results on the LSRV road dataset

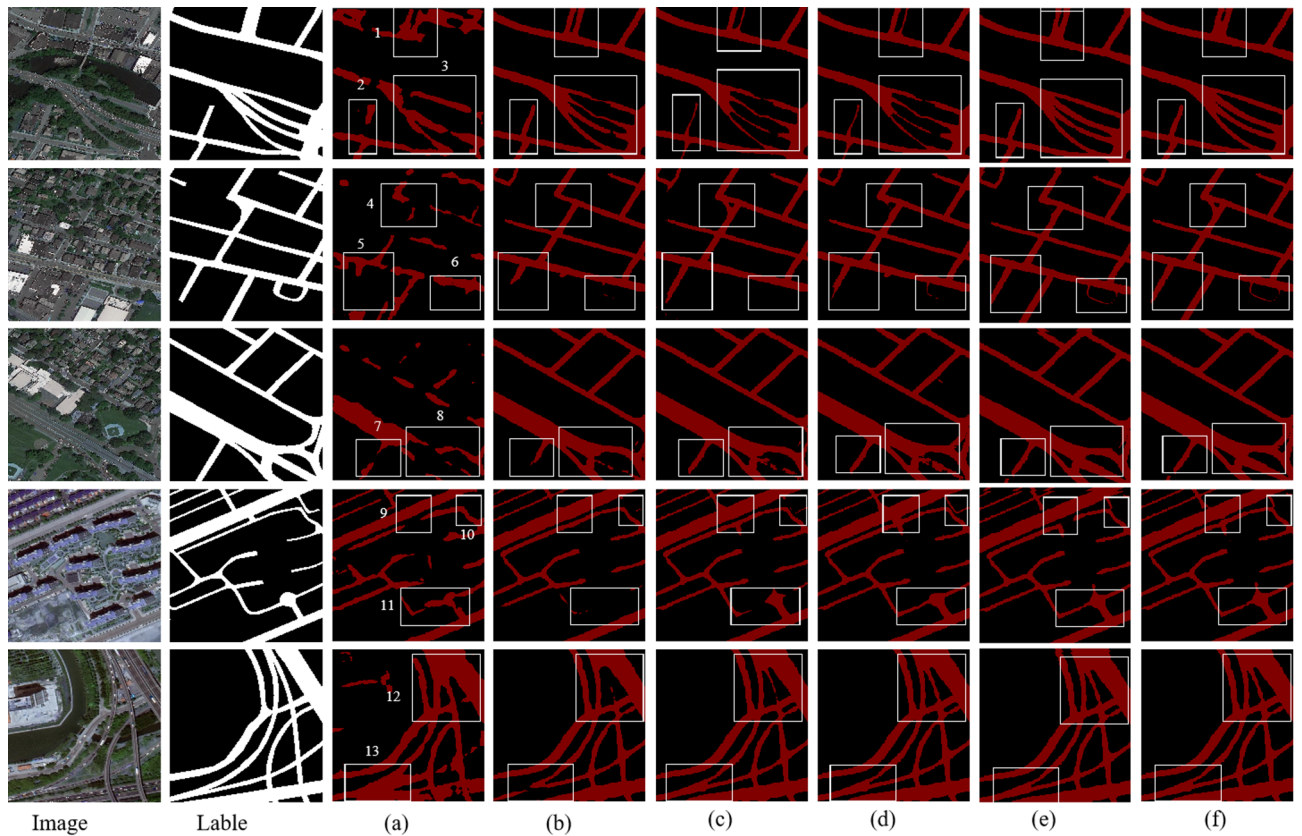
In this stage of the study, we conducted an in-depth analysis of the model's performance on the LSRV road dataset. The LSRV dataset includes urban roads from Birmingham, Boston, and Shanghai, providing a good opportunity to assess the model's ability to identify roads within urban environments. Figure 12 presents the segmentation results for these areas.

Image1 and Image2 are remote sensing images of the city of Birmingham, where vegetation is relatively abundant. DenseUNet (c), MDCGANet (d), and SCSM (e) perform well, effectively identifying roads even with significant vegetation occlusion. This mainly benefits from the multi-scale context-aware module of MDCGANet (d) and the scene coupling mechanism of SCSM(e), which can effectively capture the correlations between road segments divided by vegetation. Meanwhile, the dense connection structure of DenseUNet (c) enhances the efficiency of feature propagation.

Image3 and Image4 are remote sensing images from Boston, where the roads are complexly interwoven. In these images, the roads are narrow and winding within the urban buildings, making them difficult to detect. Even the manually labeled images do not annotate these roads. In this case, the performance of the UNet (a) network is noticeably inferior to the previous models, with most internal roads not being identified. KanUNet (b), DenseUNet (c), and MDCGANet (d) successfully detect most of the roads, but the connections between roads are still incomplete, with many fragmented sections, because The dense connections in DenseUNet (c)

Scheme	Network	Precision (%)	Recall (%)	IoU (%)	F1-Score (%)
(a)	UNet	88.81	75.26	77.80	81.64
(b)	KanUNet	89.75	77.73	76.03	83.26
(c)	DenseUNet	91.32	80.02	81.75	85.32
(d)	MDCGANet	91.68	82.96	77.72	87.12
(e)	SCSM	92.03	81.35	80.41	86.51
(f)	RISENet	<b>92.24</b>	<b>85.72</b>	<b>82.77</b>	<b>88.86</b>

**Table 2.** Quantitative analysis results of five models on the deepglobe road dataset. Significant values are in [bold].



**Fig. 12.** Experimental results on the LSRV road dataset. We selected 12 regions for display using white boxes. In the figure, Image represents the original remote sensing road image, Lable represents the manual labeling label, (a) represents the UNet extraction.

Scheme	Network	Precision (%)	Rcall (%)	IoU (%)	F1-Score (%)
(a)	UNet	85.40	73.07	73.34	78.58
(b)	KanUNet	86.58	75.56	77.22	80.70
(c)	DenseUNet	86.84	76.63	<b>78.33</b>	81.34
(d)	MDCGANet	87.32	77.14	75.42	81.96
(e)	SCSM	87.96	77.51	77.19	82.42
(f)	RISENet	<b>88.18</b>	<b>77.85</b>	76.25	<b>82.70</b>

**Table 3.** Quantitative analysis results of the five models on the LSRV road dataset. Significant values are in [bold].

lead to a relatively large number of parameters. In narrow road scenarios, due to the interference of redundant features, the connection of fragmented road sections may be incomplete. In MADCGANet (d), the training of the GAN is unstable, and it is prone to generate false road segments, resulting in fluctuations in the accuracy rate. RISENet (f), however, is able to identify even the narrow roads with high accuracy, producing relatively complete road maps compared to the other models, showcasing its excellent performance.

Image5 uses a remote sensing image of Shanghai, where the city roads feature complex interchanges and bridges, resulting in inevitable occlusion between the bridges and roads. DenseUNet (c) and SCSM (e) networks identify discontinuous road segments in areas with vehicle occlusion, showing some recognition errors. RISENet (f), on the other hand, identifies the roads almost identically to the manual labels, with no missed or incorrect detections.

The quantitative analysis of the LSRV road dataset is provided in the following Table 3.

In the LSRV dataset, the accuracy of various models is lower compared to the DeepGlobe and Massachusetts datasets, which may be due to the excessive road detail in urban areas of the LSRV dataset, making it difficult for the models to accurately identify roads. KanUNet achieves an accuracy of 87.38% and 89.75% in the Massachusetts and DeepGlobe datasets, respectively, but reaches 86.58% in the LSRV dataset, the compact context framework of KanUNet performs stably in moderately complex scenes, but its feature discrimination

ability is limited when facing extreme details. This suggests that KanUNet performs better in the LSRV dataset than expected, indicating its adaptability to complex scenes. RISENet's F1-score is 87.98% and 88.86% in the Massachusetts and DeepGlobe datasets, respectively, while it is 82.70% in the LSRV dataset. Although RISENet's performance on the LSRV dataset is not as high as on the other two datasets, its F1-score still outperforms other models. Overall, despite fluctuations in performance across the LSRV dataset, RISENet maintains a high F1-score, demonstrating its strong generalization ability across different datasets.

Through experimental comparisons of three different types of road datasets, the results clearly demonstrate that our developed RISENET model has exhibited excellent performance in various complex road scenarios. Whether it is the winding rural roads, the mountainous roads with complex terrains, or the urban roads with complex layouts, the RISENET model has performed outstandingly. Especially in the urban road scenarios, although urban roads are generally characterized by density and narrowness, making the recognition extremely difficult, the RISENet model, with its powerful global perception ability and precise local attention advantages bestowed by its unique design, can accurately capture road features and effectively overcome recognition challenges, the dual-branch encoder processes spatial details and semantic context separately, and the hybrid decoder restores the clarity of the road boundaries through the feature dilation operation. Compared with other models, the performance advantages of the RISENET model in urban road scenarios are extremely remarkable, and its recognition accuracy and stability far exceed those of similar models, fully demonstrating its strong adaptability and excellent performance in complex urban road environments.

### Ablation experiment

In evaluating the RISENet model, we conducted ablation experiments on the Massachusetts, DeepGlobe, and LSRV road datasets to assess the contribution of each module to performance. The ablation experiment consists of the following components:

- (a) DoubleC + Conv: This is a basic encoder-decoder model that does not include any enhancement modules. The purpose of this experiment is to demonstrate the performance of the basic model as a reference for other enhanced models.
- (b) DoubleC + Conv + MDW + FC: In this experiment, we added the MDW module and the FC module to the basic DoubleC + Conv model. This setup evaluates the impact of the feature fusion module group on the overall model performance.
- (c) DoubleC + Conv + MDW + FC + MCSA: In this configuration, the network is similar to the previous one, but without any attention mechanism in the bridge connection. We introduced the innovative MCSA module in the bridge connection to enhance the model's ability to focus on key features.
- (d) DoubleC + Conv + MCSA + FDA: Here, we placed the innovative FDA module in the decoder and combined it with MCSA. We expect this modification to improve the model's accuracy in road identification.
- (e) RISENet: This is our fully developed model, incorporating all modules into the base network. We expect this model to achieve the best performance in the ablation experiments.

By incrementally adding modules to the base DoubleC + Conv network, each model exhibits varying degrees of enhancement. This step-by-step integration of our innovative modules allows for a precise analysis of the contribution of each component to the model's performance. To validate the effectiveness of our model in practical applications, we conducted both qualitative and quantitative analyses on representative remote sensing images from the Massachusetts, DeepGlobe, and LSRV road datasets. The quantitative results are shown in Tables 4, 5 and 6.

After completing the above ablation quantitative experiments, we further analyzed the interactions among various modules. In the encoding stage, the DoubleC and MDW modules work together to form a multi-scale feature extraction strategy. DoubleC, as a basic convolutional unit, initially extracts the local edge information of the image, while MDW further captures road features at different scales. After both process the input image of the same layer in parallel, their features are sent to the FC module for fusion. The role of the FC module is to integrate features from different pathways, ensuring that information at different scales can be utilized more effectively in subsequent layers without information loss due to feature dispersion. On the three datasets, the average accuracy increased by 9.19% after adding the feature fusion module group. Between the encoder and the decoder, the MCSA module plays a bridging role in the model. Its core is to dynamically allocate attention weights between different spaces and channels. This enables the bridging layer to enhance feature selection ability while transmitting information, effectively reducing the interference of irrelevant information, and thus improving the

D + C	MDW	FC	MCSA	FDA	Massachusetts			
					P (%)	R (%)	IoU (%)	F1 (%)
√					86.65	77.53	74.64	81.88
√	√	√			86.81	78.21	64.93	82.22
√	√	√	√		88.52	84.67	75.82	86.54
√			√	√	87.91	82.78	65.26	85.22
√	√	√	√	√	<b>90.04</b>	<b>86.03</b>	<b>82.01</b>	<b>87.98</b>

**Table 4.** Results of five ablation experiments designed on Massachusetts road dataset. Significant values are in [bold].

D+C	MDW	FC	MCSA	FDA	DeepGlobe			
					P (%)	R (%)	IoU (%)	F1 (%)
√					88.37	73.90	70.15	80.44
√	√	√			88.54	75.66	73.10	81.58
√	√	√	√		89.41	81.80	78.54	85.46
√			√	√	89.09	79.44	78.15	83.96
√	√	√	√	√	<b>92.24</b>	<b>85.72</b>	<b>82.77</b>	<b>88.86</b>

**Table 5.** Results of five ablation experiments designed on deepglobe road datasets. Significant values are in [bold].

D+C	MDW	FC	MCSA	FDA	LSRV			
					P (%)	R (%)	IoU (%)	F1 (%)
√					84.71	68.03	61.54	75.44
√	√	√			85.43	68.54	63.47	76.08
√	√	√	√		87.21	77.53	74.97	81.98
√			√	√	86.37	70.72	65.69	77.74
√	√	√	√	√	<b>88.18</b>	<b>77.85</b>	<b>76.25</b>	<b>82.72</b>

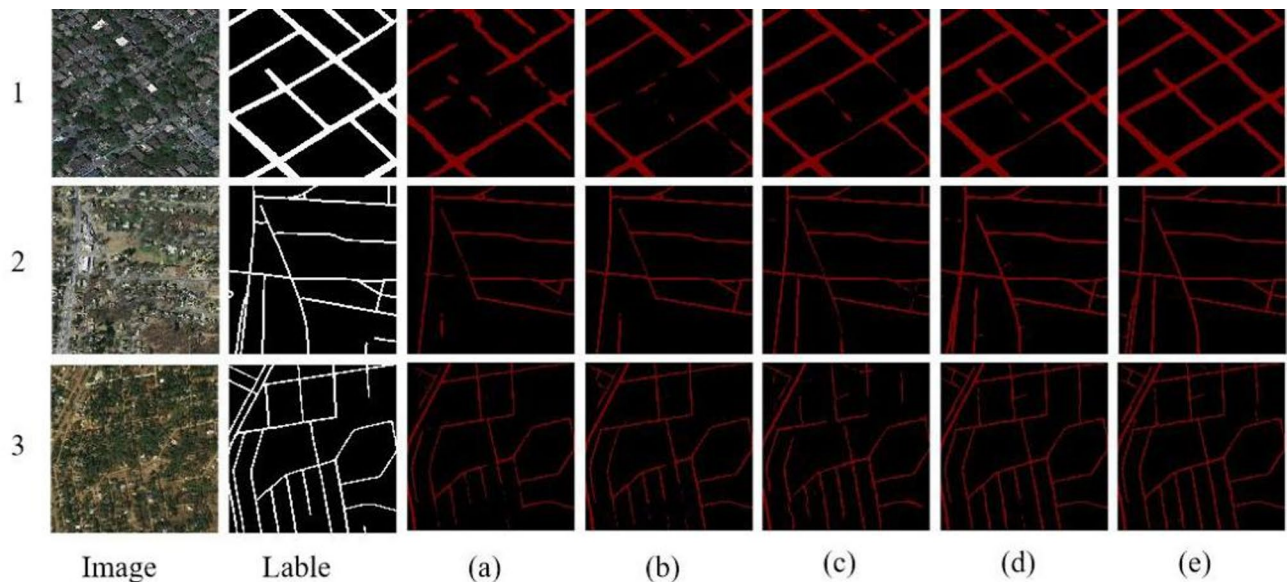
**Table 6.** Results of five ablation experiments designed on LSRV road dataset. Significant values are in [bold].

extraction accuracy of road features. Compared with only adding the feature fusion module group, the accuracy increased by up to 1.78% after adding the MCSA module. In the decoding stage, the introduction of the FDA module greatly enhances the integration ability of multi-scale features. The FDA adopts the HDC (Hybrid Dilated Convolution) technology, expanding the receptive field by combining different dilation rates. The input of the FDA comes from two sources: one is the four-layer feature maps output by MCSA, and the other is the feature map output by the previous-layer decoder. The two are first fused through a concatenation operation, and then deep-level semantic information is extracted through HDC, and bilinear interpolation is used for upsampling, so that the decoding result can more accurately match the size of the original image. This function of the FDA ensures that the decoder can fully utilize the multi-scale features extracted by the encoder, while enhancing the ability to accurately locate road boundaries. In the entire model, DoubleC and MDW complete local and global feature extraction in the encoding stage and are fused through FC. MCSA builds an effective feature bridge between encoding and decoding. The FDA further improves the utilization rate of multi-scale features in the decoding stage, ensuring that the final road extraction result has stronger integrity and edge detail expressiveness. This synergistic effect among modules enables RISENet to effectively reduce problems such as breaks and blurs when dealing with complex road environments, and ultimately demonstrates excellent performance on multiple datasets. The qualitative results are shown in Fig. 13.

In the road extraction task, we selected three representative road scenes: a densely built-up area (1), a sparsely built-up area (2), and a densely vegetated area (3). Our baseline model performs adequately in road extraction but tends to suffer from discontinuities, particularly when handling wide roads, as observed in scenes (1) and (2). Experimental results indicate that combining the baseline model with the feature fusion module group leads to a modest improvement in accuracy, especially in filling the gaps in wide roads. The combination of D+C+MDW+FC+MCSA and D+C+FDA+MCSA produces more favorable road segmentation images, eliminating the problem of discontinuities in wide roads, making the extraction results more coherent, and improving the precision of road edge predictions. Notably, in scenes (2) and (3), RISENet demonstrates its advantage in road prediction, producing clearer predictions with smoother and more regular edge predictions, even achieving precise segmentation at road corner edges. Finally, qualitative experimental results confirm that the combination of MDW, FC, MCSA, and FDA modules effectively learns road boundary features and accurately predicts road edges, with results closer to the actual situation.

## Conclusion

To address the challenges of road segmentation in remote sensing images under complex scenarios, such as discontinuity, incompleteness, and blurry edges, this study proposes a high-precision road extraction model, RISENet. This model adopts an encoder-decoder architecture, with MDW modules and feature fusion modules added to the encoder to enhance feature perception capabilities. In the skip connections, a Multi-layer Dynamic Spatial Channel Fusion Attention mechanism is introduced, which strengthens key features by differentiating weight allocation and reduces the impact of non-key features, allowing the model to focus more on the road parts of the image. Finally, the decoder consists of FDA modules and convolution modules. The FDA module extracts multi-scale road features by gradually expanding the receptive field, optimizing feature fusion, and enhancing the model's ability to comprehensively capture information at different scales in complex scenes. Experimental results on the Massachusetts road dataset, DeepGlobe road dataset, and LSRV dataset demonstrate that the improvements in the RISENet model structure and the innovative attention mechanism



**Fig. 13.** Ablation experimental performance results of five schemes on Massachusetts road dataset, DeepGlobe road dataset and LSRV road dataset. Where 1,2,3 represent remote sensing images from Massachusetts road dataset, DeepGlobe road dataset, and LSRV road.

are significantly effective, widespread, and robust. Looking ahead, we will continuously carry out comprehensive and in-depth optimization of the RISENet model. On the one hand, we focus on promoting lightweight design while improving extraction accuracy. Specifically, channel pruning techniques will be employed to eliminate redundant convolutional kernels, and the dynamic threshold method will be utilized to retain the channels that play a crucial role in feature extraction. By doing so, we aim to reduce computational complexity and resource consumption, thereby ensuring that the model can operate efficiently in various different hardware environments. On this basis, we plan to introduce Fourier transform technology into RISENet and process the feature images in the frequency domain. Through designing a learnable frequency domain filter bank and dynamically adjusting the weights of each frequency component with the help of the attention mechanism, we will focus on enhancing the medium and high-frequency components that represent the edge features of roads. By fully exploiting the detailed features contained in the frequency information, we can further improve the accuracy of road extraction, achieving a dual improvement in both precision and efficiency. On the other hand, we will further expand the application scope of RISENet in the field of remote sensing. We plan to apply it to tasks such as building extraction and vegetation extraction. To this end, we plan to construct a multi-scale test set with a resolution ranging from 0.3 to 2 m, covering 10 different geomorphic types such as cities, farmlands, and mountains, so as to ensure the robustness of the model in a wide range of scenarios. Meanwhile, we will also establish a multi-dimensional evaluation system. In addition to using the conventional intersection over union (IoU) index, we will add new indexes, namely the edge continuity index and the shape preservation degree index, to quantitatively evaluate the stability of the model in complex scenarios, thus providing a solid guarantee for the extended application of RISENet in multiple fields.

### Data availability

The datasets used for generating the satellite imagery in this study are as follows: Figures 1, 2, 4 and 5, and 11 were generated using the DeepGlobe Road Dataset, processed with Python 3.8 and PyTorch 1.8, and the dataset can be requested through the link <https://hyper.ai/cn/datasets/34270>, which is publicly available under the CC BY 4.0 license; Figure 10 was generated using the Massachusetts Road Dataset, processed with Python 3.8 and PyTorch 1.8, and the dataset is accessible via <http://web.mit.edu/torralba/www/scene/paris6/road-detection.html>, also under the CC BY 4.0 license; Figure 12 was generated using the LSRV Road Dataset, processed with Python 3.8 and PyTorch 1.8, and the dataset is available at [http://rsidea.wlu.edu.cn/resource\\_LSRV\\_sharing.htm](http://rsidea.wlu.edu.cn/resource_LSRV_sharing.htm), permitted for academic use under the CC BY 4.0 license, with explicit permission obtained from the copyright holders; and Figure 13 consists of composite images derived from the aforementioned datasets, with all modifications performed under CC BY 4.0 and permission from the respective copyright holders. The code used in this study will be released via GitHub and is available upon request from the corresponding author.

Received: 19 December 2024; Accepted: 12 May 2025

Published online: 21 May 2025

### References

1. Chen, C., Seff, A., Kornhauser, A. & Xiao, J. DeepDriving learning affordance for direct perception in autonomous driving. In *IEEE International Conference on Computer Vision (ICCV) 2722–2730* (IEEE, 2015). <https://doi.org/10.1109/ICCV.2015.312>.

2. Traffic Management as a Smart City Solution. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* 1–6 (IEEE, 2020). <https://doi.org/10.1109/CONECCT50063.2020.9198588>.
3. Szántó, M. et al. Trajectory planning of automated vehicles using Real-Time map updates. *IEEE Access*. **11**, 67468–67481 (2023).
4. Dai, R., Xu, S., Gu, Q., Ji, C. & Liu, K. Hybrid spatio-temporal graph convolutional network: improving traffic prediction with navigation data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 3074–3082 (ACM, 2020). <https://doi.org/10.1145/3394486.3403358>.
5. Hu, W. C., Wu, H. T., Cho, H. H. & Tseng, F. H. O. *ptimal Route Planning System for Logistics Vehicles Based on Artificial Intelligence* (Springer, 2022).
6. Anil Chougule, M. & Mashalkar, A. S. A comprehensive review of agriculture irrigation using artificial intelligence for crop production. In *Computational Intelligence in Manufacturing* 187–200 (Elsevier, 2022). <https://doi.org/10.1016/B978-0-323-91854-1.00002-9>.
7. Tong, Z., Ye, F., Yan, M., Liu, H. & Basodi a survey on algorithms for intelligent computing and smart City applications. *Big Data Min. Anal.* **4**, 155–172 (2021).
8. Yu, D. & Fang, C. Urban remote sensing with Spatial big data: a review and renewed perspective of urban studies in recent decades. *Remote Sens.* **15**, 1307 (2023).
9. Lian, R., Wang, W., Mustafa, N. & Huang, L. Road extraction methods in High-Resolution remote sensing images: a comprehensive review. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **13**, 5489–5507 (2020).
10. Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S. & Alamri, A. Deep learning approaches applied to remote sensing datasets for road extraction: A State-Of-The-Art review. *Remote Sens.* **12**, 1444 (2020).
11. Wei, Y., Zhang, K. & Ji, S. Simultaneous road surface and centerline extraction from Large-Scale remote sensing images using CNN-Based segmentation and tracing. *IEEE Trans. Geosci. Remote Sens.* **58**, 8919–8931 (2020).
12. Liu, R. et al. A review of deep Learning-Based methods for road extraction from High-Resolution remote sensing images. *Remote Sens.* **16**, 2056 (2024).
13. Pruthi, J. & Dhingra, S. A. Review of research on road feature extraction through remote sensing images based on deep learning algorithms. In *3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)* 1–5 (IEEE, 2023). <https://doi.org/10.1109/CISCT57197.2023.10351299>
14. Liu, X. et al. RoadFormer: road extraction using a Swin transformer combined with a Spatial and channel separable Convolution. *Remote Sens.* **15**, 1049 (2023).
15. Khan, M. J. & Singh, P. P. Advanced road extraction using CNN-based U-Net model and satellite imagery. *e-Prime - Adv. Electr. Eng. Electron. Energy*. **5**, 100244 (2023).
16. Xu, Q., Long, C., Yu, L. & Zhang, C. Road extraction with satellite images and partial road maps. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–14 (2023).
17. Gong, Z., Xu, L., Tian, Z., Bao, J. & Ming, D. Road network extraction and vectorization of remote sensing images based on deep learning. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)* 303–307 (IEEE, 2020). <https://doi.org/10.1109/ITOEC49072.2020.9141903>.
18. Qiu, L., Yu, D., Zhang, C. & Zhang, X. A Semantics-Geometry framework for road extraction from remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5 (2023).
19. Lian, R. & Huang, L. DeepWindow Sliding window based on deep learning for road extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **13**, 1905–1916 (2020).
20. Yang, Z., Zhou, D., Yang, Y., Zhang, J. & Chen, Z. Road extraction from satellite imagery by road context and Full-Stage feature. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5 (2023).
21. Bobba, S. Leveraging Pre-trained deep learning models for remote sensing image classification: A case study with ResNet50 and EfficientNet. *AJSET* **9**, 150–162 (2024).
22. Lv, L. et al. A deep learning network for individual tree segmentation in UAV images with a coupled CSPNet and attention mechanism. *Remote Sens.* **15**, 4420 (2023).
23. Dai, L., Zhang, G. & Zhang, R. RADANet Road augmented deformable attention network for road extraction from complex High-Resolution Remote-Sensing images. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–13 (2023).
24. Wang, Q. et al. IEEE, Seattle, WA, USA., ECA-Net: efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11531–11539 (2020). <https://doi.org/10.1109/CVPR42600.2020.01155>.
25. Cao, S. et al. BEMRF-Net: boundary enhancement and multiscale refinement fusion for Building extraction from remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **17**, 16342–16358 (2024).
26. Xu, W. et al. Building height extraction from High-Resolution Single-View remote sensing images using shadow and side information. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **17**, 6514–6528 (2024).
27. Xie, Y. et al. Localization, balance, and affinity: a stronger multifaceted collaborative salient object detector in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **63**, 1–17 (2025).
28. Zhu, J. et al. A cross-view intelligent person search method based on multi-feature constraints. *Int. J. Digit. Earth.* **17**, 2346259 (2024).
29. Hong, Z., Ming, D., Zhou, K., Guo, Y. & Lu, T. Road extraction from a high Spatial resolution remote sensing image based on richer convolutional features. *IEEE Access.* **6**, 46988–47000 (2018).
30. Vaswani, A. et al. Attention is all you need (2022).
31. Engel, N., Belagiannis, V. & Dietmayer, K. Point transformer. *IEEE Access.* **9**, 134826–134840 (2021).
32. Wang, R., Cai, M., Xia, Z. & Zhou, Z. Remote sensing image road segmentation method integrating CNN-Transformer and UNet. *IEEE Access.* **11**, 144446–144455 (2023).
33. Beltagy, I., Peters, M. E. & Cohan, A. Longformer the long-document transformer (2020). <https://doi.org/10.48550/arXiv.2004.05150>.
34. Child, R., Gray, S., Radford, A. & Sutskever, I. Generating long sequences with sparse transformers (2019). <https://doi.org/10.48550/arXiv.1904.10509>.
35. Chen, W. et al. A simple single-scale vision transformer for object localization and instance segmentation (2022). <https://doi.org/10.48550/arXiv.2112.09747>.
36. Zhang, X. et al. HIVIT: a simpler and more efficient design of hierarchical vision transformer (2023).
37. Fu, Z. *Vision Transformer Vit and its Derivatives* (2022). <https://doi.org/10.48550/arXiv.2205.11239>.
38. Yin, H. et al. IEEE, New Orleans, LA, USA., A-ViT: adaptive tokens for efficient vision transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10799–10808 (2022). <https://doi.org/10.1109/CVPR52688.2022.01054>.
39. Zhou, D. et al. *DeepViT: Towards Deeper Vision Transformer* (2021). <https://doi.org/10.48550/arXiv.2103.11886>.
40. Lu, W., Shi, X. & Lu, Z. A new two-step road extraction method in high resolution remote sensing images. *PLoS ONE.* **19**, e0305933 (2024).
41. Ronneberger, O., Fischer, P. & Brox, T. *U-Net Convolutional Networks for Biomedical Image Segmentation* (2015). <https://doi.org/10.48550/arXiv.1505.04597>.
42. Xin, J., Zhang, X., Zhang, Z. & Fang, W. Road extraction of High-Resolution remote sensing images derived from DenseUNet. *Remote Sens.* **11**, 2499 (2019).

43. Niu, P. et al. MDCGA-Net: multiscale direction Context-Aware network with global attention for Building extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **17**, 8461–8476 (2024).
44. Ma, X. et al. *A Novel Scene Coupling Semantic Mask Network for Remote Sensing Image Segmentation* (2025). <https://doi.org/10.48550/arXiv.2501.13130>.

### Author contributions

H.B. wrote the main manuscript text and Z. H. and Y.G prepared figures 1-20, and C.R. was responsible for the overall supervision of the work. All authors reviewed the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to C.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025