Scientific Reports

https://doi.org/10.1038/s41598-026-35723-y

Article in Press

Multi-feature enhancement fusion network for remote sensing image semantic segmentation

Received: 7 October 2025

Accepted: 7 January 2026

Published online: 11 January 2026

Cite this article as: Zhang W., Yang W., Yin Y. *et al.* Multi-feature enhancement fusion network for remote sensing image semantic segmentation. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-35723-y

Wansong Zhang, Wenzhong Yang, Yabo Yin, Danny Chen, Xianfeng Wang & Hu Zhao

Multi-Feature Enhancement Fusion Network for Remote Sensing Image Semantic Segmentation

Wansong Zhang[1,2], Wenzhong Yang[1,2*], Yabo Yin[1,2*], Danny Chen[1,2], Xianfeng Wang[1,2], and Hu Zhao[1,2]

[1]Xinjiang University,School of Computer Science and Technology (School of Cyberspace Security), Urumqi, 830046, China
[2]Xinjiang University, Xinjiang Key Laboratory of Multilingual Information Technology, Urumqi, 830046, China
[*]Corresponding authors: yangwenzhong@xju.edu.cn, yinyabo@xju.edu.cn

ABSTRACT

Semantic segmentation of remote sensing images has important application value in fields such as farmland anomaly detection and urban planning. However, the low-level features extracted by deep neural network models retain rich spatial detail information while introducing redundancy and noise. The significant differences in the semantic level and spatial distribution of high-level and low-level features pose challenges to their effective fusion. To this end, we propose a Multi-Feature Enhancement Fusion Network that improves local feature expression and global semantic modelling ability by fusing edge information and semantic information. The Edge Enhancement Module used traditional edge detection operators to enhance the details of edge features. The Multi-Feature Fusion Module effectively integrates semantic and edge features to enhance the ability to express fine-grained information. The Local-Global Feature Enhancement Module hierarchically establishes local details and global context information, and the Multi-Level Fusion segmentation head integrates the features of different levels to utilise both shallow spatial details and deep semantic information fully. Following this, our extensive experiments on three publicly available datasets demonstrate that the proposed model outperforms state-of-the-art methods. The code will be published on: https://github.com/zwsbh/MFEF.

Keywords: Remote sensing image, State space model, Multi-Feature fusion, Edge enhancement, Semantic segmentation

Introduction

Semantic segmentation holds a core position in remote sensing image processing[1]. Its task is to perform pixel-level classification on the original image, thereby extracting semantic information with clear categories. With the rapid development of unmanned aerial vehicles and aerospace technology, the acquisition of high-resolution remote sensing images has become more convenient than ever before[2,3]. The application scenarios of semantic segmentation have been expanded to multiple fields such as urban planning and farmland anomaly monitoring[4–7] . In these applications, how to precisely extract semantic information from the original images, that is, to assign clear category labels to each pixel such as abnormal types of vegetation, buildings or farmland, is not only a key link in remote sensing image processing, but also directly determines the accuracy and reliability of subsequent geographic information analysis[8,9].

As shown in Figure 1, compared with natural scene images, remote sensing images face greater challenges. On the one hand, different ground objects often show a high degree of similarity, while the same ground objects show significant differences due to differences in texture, material and imaging conditions[10–12]. On the other hand, the perspective change and occlusion in the process of aerial photography further aggravate the difficulty of recognition. These factors make the traditional methods that rely on manual work difficult to adapt to complex and diverse ground objects distribution, and the accuracy and generalization ability are limited[13,14].

In recent years, deep learning has driven significant progress in semantic segmentation. Convolutional neural networks perform well in local feature extraction. FCN achieves end-to-end pixel-wise prediction[15]. Encoder-decoder structures such as U-Net fuse shallow details and deep semantics through multi-level skip connections, and have achieved outstanding results in medical and remote sensing image segmentation. However, CNNs are limited to a fixed receptive field, which makes it difficult to model long-distance dependencies and global context. In order to break through this limitation, researchers introduce the Vision Transformer to realise global modelling with the help of the self-attention mechanism. For multi-scale expression, Swin Transformer achieves a balance between performance and efficiency by layering and shifting Windows and becomes the mainstream scheme. However, when dealing with high-resolution images, the computational and storage overhead is still huge, and there are shortcomings in the modelling of shallow boundaries and local details, which easily lead to limited boundary

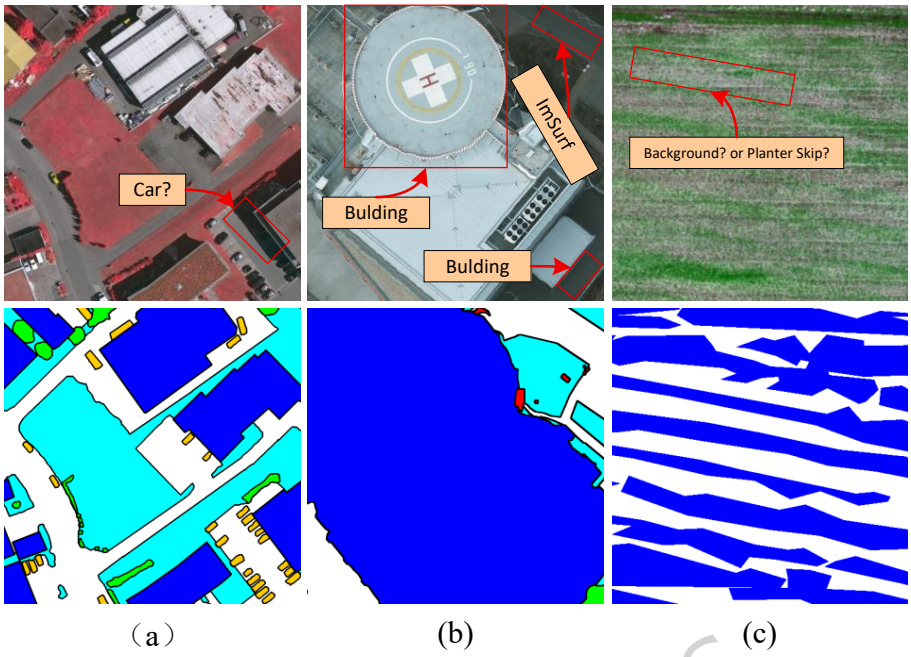

Figure 1. Some examples are challenging in semantic segmentation of remote sensing images. (a) Vehicles are occluded by buildings under certain light angles. (b) The same category has different textures and shapes, and the textures are similar across categories. (c) High similarity between different categories.

clarity and semantic consistency of segmentation results. Therefore, how to effectively enhance the shallow spatial details and boundary features while maintaining the global context modelling ability, and achieve a dynamic balance between semantic information and edge structure, is a challenge worthy of further research.

To address the above issues, we propose a Multi-Feature Enhancement Fusion (MFEF) network, as shown in Figure 2(a). We design an Edge Enhancement Module (EEM) to strengthen the edge information in the underlying feature map. Subsequently, edge and semantic features are introduced into the Multi-Feature Fusion Module (MFFM) to achieve effective integration of the two, thereby enhancing the model's ability to represent fine-grained structures. On this basis, the Local-Global Feature Enhancement module (LG-FEM) is further introduced. Through the division of feature sub-blocks and the establishment of long-distance dependencies, global and local context information is effectively captured, enhancing the global consistency of feature representation and the ability to restore local details. Ultimately, the Multi-Level are fused to obtain the segmentation result. This research has made the following main contributions.

- We designed an LG-FEM , in the stage following multi-feature fusion, we enhanced the recovery ability of local details in the model through the sub-block segmentation strategy to improve the modelling effect of the global context.
- We designed the EEM , where traditional edge detection operators are utilised to pre-enhance the boundary information of ground objects from the low-level feature map, to improve the accuracy and continuity of the target contour.
- We proposed the MFFM, which achieves an effective balance between edge details and semantic understanding by cross-fusing the underlying edge features with semantic context information.
- We conduct extensive experiments on three publicly available benchmark datasets to verify the effectiveness of MFEF-UNet.

Related work

CNN-Based Remote Sensing Image Semantic Segmentation

As the foundation work of semantic segmentation, FCN realizes the end-to-end pixel-level prediction for the first time, and opens up a new direction for image segmentation using CNNs[15]. Subsequently, a large number of researches have been devoted

to alleviating the problem of detail loss caused by downsampling, and improving segmentation performance by expanding receptive field and multi-scale feature fusion[16]. For example, the Deeplab family of methods introduces dilated convolution to effectively expand the receptive field of the convolution kernel[17]. PSPNet fuses multi-scale context information to effectively capture and express features at different scales[18]. U-Net effectively recovers spatial details through encoder-decoder structure and skip connection, which is widely used in medical and remote sensing image segmentation[19].

However, the limitation of receptive field makes it difficult for these methods to fully establish global context information. ABCNet uses bilateral contextual attention mechanism to enhance global semantic modeling[20], and SFFNet uses pyramid pooling structure to extract multi-scale features[21]. Although these methods expand the receptive field, their ability to capture global context information is still limited by the inherent characteristics of convolution operation[22].

Tansformer-Based Remote Sensing Image Semantic Segmentation

The success of Transformers in natural language processing has driven their widespread use in computer vision[23–25]. Vision Transformer (ViT) is the first application of Transformers to vision tasks. The self-attention mechanism in vit has the natural advantage of global modeling, which is able to effectively capture long-distance dependencies in semantic segmentation tasks[26]. Then, the Swin Transformer reduces the computational overhead through the hierarchical structure and sliding window self-attention[27]. It is applied in SegFormer and Segmenter methods[28,29]. In addition, the TransUNet hybrid architecture combines the local detail extraction of CNN and the global modeling of Transformer, showing superior performance in remote sensing segmentation[30]. LSRFormer combines convolutional networks with efficient long-short range transformers to supplement the global semantics[31] after each level of CNN. Although Transformer-based methods have significant advantages in global semantic capture and multi-scale feature fusion, their high computational cost and training difficulty still restrict large-scale applications. Therefore, how to design lightweight Transformer architectures that balance performance and efficiency has become a key direction of current research[32].

Mamba-Based Remote Sensing Image Semantic Segmentation

Recently, Mamba architecture has been introduced into the field of computer vision as a new sequence modeling method. Its core is based on the State Space Model (SSM), which can realize remote dependency modeling while maintaining linear computational complexity[33]. Compared with the self-attention mechanism of Transformer, Mamba has a lower memory footprint and faster inference speed on long sequence processing, so it has potential advantages in large-scale high-resolution image segmentation tasks[34,35].

In semantic segmentation tasks, Mamba structure can effectively model spatial-temporal features through state update and input mapping, and further enhance local context awareness by combining convolution operation[36,37]. It shows high performance and application potential in high-resolution scenes such as remote sensing and medical imaging. The UMFormer model combines Mamba module with convolution to achieve a balance between global semantics and local detail modeling[38]. Mamba's efficiency in multi-scale feature fusion and long-range dependency modeling. RSMamba architecture achieves global modeling and efficient classification of two-dimensional remote sensing images through state space Model (SSM) and introducing dynamic multi-path activation mechanism[39].

Methods

In this section, we provide an overview of the architecture of the proposed MFEF-UNet and further elaborate on its core modules.

MFEF-UNet Framework

This is shown in Figure 2(a). The MFEF-UNet network consists of an encoder and a decoder. The encoder part uses the pre-trained CSWin Transformer as the backbone[40] to extract multi-scale semantic features. Through the cross-window self-attention mechanism, CSWin models long-distance dependencies in the horizontal and vertical directions while preserving local details, thereby improving the richness and robustness of feature representation. Its hierarchical structure can gradually capture and fuse semantic information at different scales, providing a solid feature foundation for downstream segmentation tasks.

In the decoder part, the MFFM is designed to replace the traditional skip connection to realize the efficient interaction and fusion of encoded features, decoded features and edge features. Then, LG-FEM is used to divide the fused feature pairs into sub-blocks to enhance the local information and take into account the recovery ability of global structure and details. Finally, the feature maps output by the multi-stage decoder were upsampled to a unified resolution and multi-fused to obtain the final segmentation result. We present our key modules in detail.

Local-Global Feature Enhancement Module

As illustrated in Figure 2(b). Unlike traditional CNN-Transformer architectures, LG-FEM leverages linear complexity VSSMS to bridge the gap between local detail capture and global context modeling. The core operating principle lies in the hierarchical

"chunk-reassemble" mechanism. Initially, we introduce a sub-block segmentation strategy in the blocking stage to solve the inherent local coherence loss in standard SSM. When the 2D feature map is expanded into a 1D sequence, spatially adjacent pixels may be mapped to distant positions in the sequence, which destroys the spatial continuity of the local neighborhood and leads to the loss of local structural information. By restricting the 2D-SSM scan to a local window, the continuity of the neighboring space is restored and the focus is on extracting fine-grained geometric textures. For the problem of global information loss across sub-blocks, in the integration stage, the sub-block features are recovered and 2D-SSM is used to interact across sub-blocks to obtain global information. On this basis, the feature enhancement module is further combined with CBAM[41] to effectively strengthen the expression of detailed features and suppress the response of redundant channels.

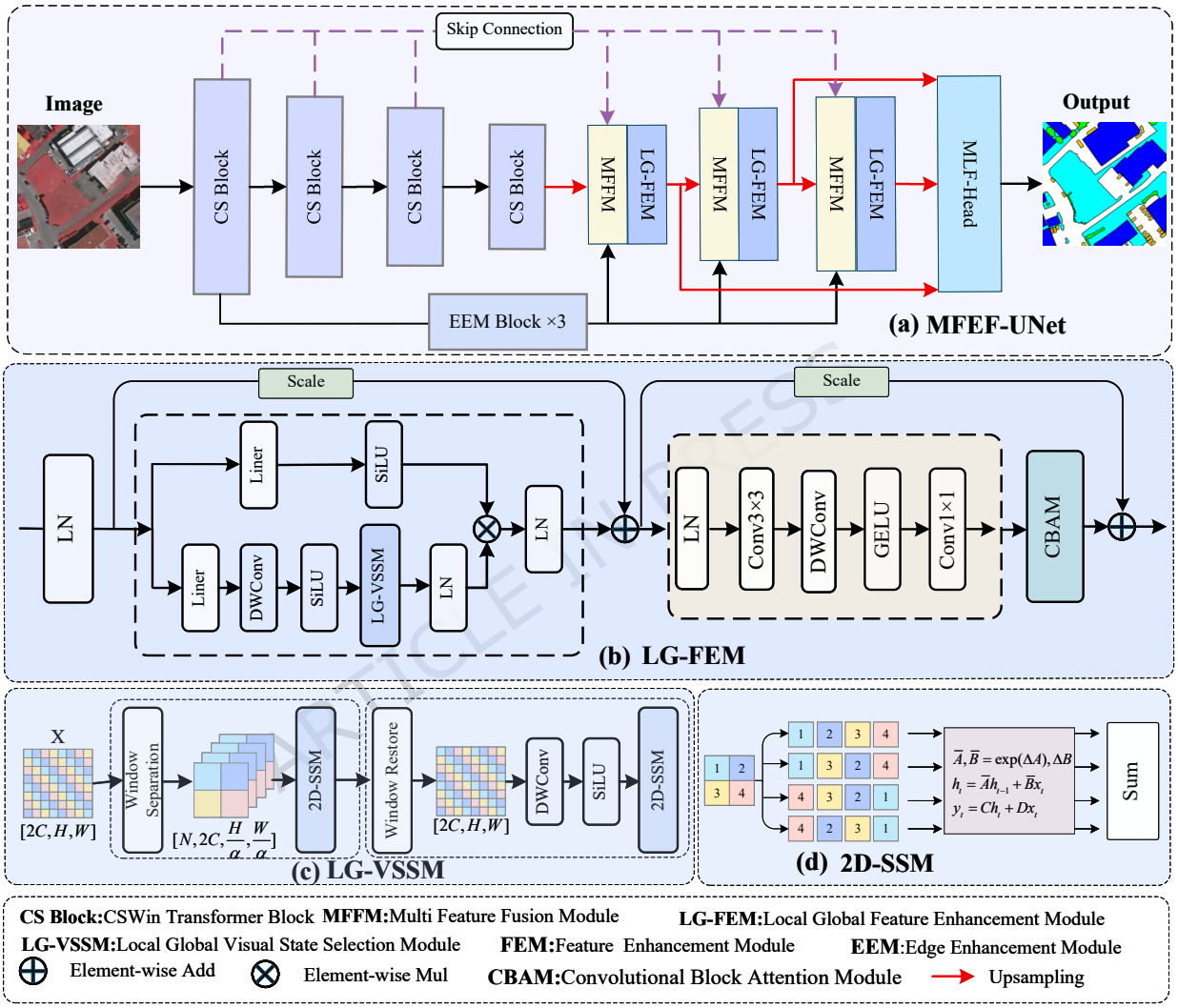

Figure 2. Overview of the MFEF-UNet model: The EEM Block module extracts edge information from feature maps. The MFFM module applies a cross-attention mechanism to the multi-source input, facilitating interaction between edge and semantic features. The LG-FEM module fully integrates global semantic information with local detailed features. The MLF-Head module fuses multi-layer, multi-scale features, thereby significantly enhancing the model's feature representation and prediction performance.

For the feature map $R \in \mathbb{R}^{C\times H\times W}$ that is input to the LG-FEM module. After the LayerNorm operation is used to normalize the channel dimension features, the standard VSSM[33] module handles the feature map in the same way, that is, the feature map is expanded along the channel dimension by linear mapping and then divided into two parts: $X \in \mathbb{R}^{2C\times H\times W}$ and $Z \in \mathbb{R}^{2C\times H\times W}$. We then pass $X \in \mathbb{R}^{2C\times H\times W}$ directly to 2D-SSM, unlike the standard VSSM operation. As shown in Figure 2(c), the feature map X is uniformly divided into multiple sub-blocks of the same size and non-overlapping to achieve local modeling of the feature map. For the selection of sub-block size, we conduct comparative experiments under different sub-block sizes.

The results show that when the window size is set to $\frac{H}{\alpha} \times \frac{W}{\alpha}$, the model achieves the optimal balance between performance and computational efficiency when α is 4. To capture the long-distance modeling ability of 2D-SSM for feature maps and compensate for the loss of local structural information during sequence flattening, all sub-blocks are treated as independent feature blocks and stacked on the batch dimension. Reshaping the feature map into a tensor of shape $F_{\mathrm{w}} \in \mathbb{R}^{N \times 2C \times \frac{H}{\alpha} \times \frac{W}{\alpha}}$. The 2D-SSM mechanism is introduced on each sub-window to better capture the relationship between the neighborhoods. Where N represents the number of sub-blocks. Under the sub-block segmentation strategy, the receptive field is expanded and the model can obtain more local information. However, since the sub-blocks are divided in a non-overlapping way, the boundary connections between sub-blocks and cross-region information interaction are inevitably limited.

While sub-block modeling enhances local granularity, it inevitably breaks global semantic continuity. To implement cross-region feature interaction to alleviate the feature fragmentation problem caused by independent sub-block modeling, we introduce a global interaction phase after sub-block recovery. Then, each sub-block is re-stitched into a complete feature map in the original order to obtain the recovered feature map, and the 2D-SSM structure in Figure 2(d) is reintroduced for global context modeling. The second 2D-SSM plays a key role in the feature interaction between sub-blocks, which acts as a "semantic bridge" connecting previously independent sub-blocks. The global cross-scan mechanism is used to capture the remote dependencies that are constrained in the partitioning phase. By integrating the 3×3convolution with nonlinear activation, this mechanism ensures that the locally enhanced features can be effectively fused into a unified global semantic space, so as to realize the collaborative modeling of local awareness and global context. After that, the dual features are multiplied and mapped back to the channel dimension of the original feature map by linear transformation as the standard VSSM processing method. The residual connection with adjustable hyperparameters is introduced to add the enhanced features and the original features element-by-element, which effectively enhances the stability and semantic consistency of the features while retaining the original information.

$$
\begin{aligned}
F_w &= \mathrm{WindowSeparation}(\mathbf{X}) \\
F_{\mathrm{s}} &= \text{2D-SSM}(\mathbf{F}_w) \\
F_r &= \mathrm{WindowRestore}(\mathbf{F}_{\mathrm{s}}) \\
F_m &= \mathrm{SiLU}\big(\mathrm{DWConv}(\mathbf{F}_r)\big) \\
Y &= \text{2D-SSM}(\mathbf{F}_{\mathrm{mid}})
\end{aligned}
\tag{1}
$$

In addition, VSSM usually introduces more hidden states to memorize very long range dependencies. The deep feature augmentation module and the hierarchical structure of CBAM[41] are used to process and promote the expressivity of different channels. After normalize the obtained features. Pass in a concatenated structure composed of depth-separable convolution and pointwise convolution. Subsequently, the CBAM attention mechanism is utilized to establish the correlation between the feature map channels and the space. This enhances the response of important features and suppresses redundant information. Finally, the output results of the module and the input features are added through the hyper-parameter residual connection to improve the discrimination and semantic consistency of the overall feature representation.

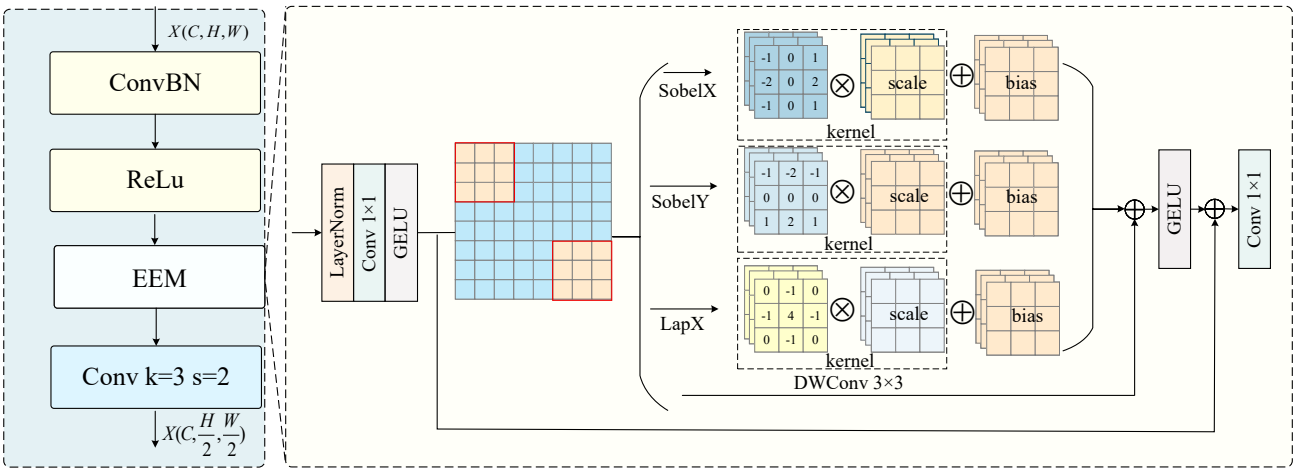

Figure 3. EEM Blocks. SobelX, SobelY and LapX represent convolutions using the Sobel-x, Sobel-y and Laplacian operators as convolution kernels.

Edge Enhancement Module

In the task of semantic segmentation of remote sensing images. Boundary regions usually contain category transitions and structural details, which are of crucial significance for precise segmentation. The establishment of boundary information can effectively enhance the sensitivity of the network to the target contour, so that the model has stronger discrimination ability at the boundary of categories. In order to effectively mine the Edge features in remote sensing images and make up for the shortcomings of the model in boundary structure perception Edge Enhancement Module is designed. As shown in Figure 3, the input to EEM is taken from the first layer features of the encoder. Compared with deep features, low-level features have higher spatial resolution and retain rich texture and structure information, which is of great significance for edge detection and detail preservation. In order to make full use of this advantage, EEM generates three layers of edge feature maps with different resolutions through multi-level processing, which provides multi-scale edge information support for the subsequent multi-feature fusion module. In EEM, the feature map is combined with traditional edge detection operators Sobel and Laplacian in the process of channel mapping to enhance the response to edge structures. For the generation of feature maps, a hierarchical structure is used to build three-layer scale feature maps step by step, where the output of the previous layer is used as the input of the next layer, and the down-sampling and channel matching operations are realized by convolution. Specifically, for the input feature $X \in \mathbb{R}^{C\times H\times W}$, we perform channel normalization and 1×1convolution to extend the channel dimension:

$$X_1 = \mathrm{GELU}\big(\mathrm{Conv}_{1\times1}(\mathrm{LN}(X))\big), X_1 \in \mathbb{R}^{\frac{C}{r}\times H\times W} \tag{2}$$

Here r denotes the channel expansion ratio.

Subsequently, a 3×3 depthwise separable convolution is introduced to enhance local spatial context modeling capability:

$$X_2 = \mathrm{DWConv}_{3\times3}(X_1) \tag{3}$$

On this basis, $\mathbf{X}_1$is further applied with multiple edge detection operators, including Sobel operator (horizontal and vertical) and Laplacian operator, to capture edge features of different directions and orders, thus establishing a close relationship between spatial details and semantic expression. The kernel of the Sobel and Laplacian operator is defined as follows.

$$K_{\text{Sobel-x}} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad K_{\text{Sobel-y}} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \quad K_{\text{Laplacian}} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{4}$$

Each operator is applied to each channel via grouped convolution, with learnable scaling factors and biases, represented as:

$$\begin{aligned} E_x &= \mathrm{Conv_g}(X_1;\ \gamma_x \cdot K_x,\ b_x) \\ E_y &= \mathrm{Conv_g}(X_1;\ \gamma_y \cdot K_y,\ b_y) \\ E_l &= \mathrm{Conv_g}(X_1;\ \gamma_l \cdot K_l,\ b_l) \end{aligned} \tag{5}$$

Here $\gamma_x, \gamma_y, \gamma_l$ are channel-wise learnable scaling factors,and b_x, b_y, b_l are the corresponding biases.

Finally, the edge enhancement feature is added to the position convolution output, and then the activation function is used for residual connection. After that, the original dimension is reduced by 1×1convolution:

$$Y = \mathrm{Conv}_{1\times1}(X_1 + \mathrm{GELU}(X_2 + E_x + E_y + E_l)) \tag{6}$$

This module introduces the prior knowledge of classical edge detection operators into the deep learning model, and realizes the adaptive adjustment of edge response through a learnable mechanism.

Multi-Feature Fusion Module

For semantic segmentation tasks that require pixel-level prediction. The rich semantic information contained in the deep features is of vital importance, while the spatial structure details and edge information contained in the shallow features are equally indispensable. To this end, we propose a Multi-Feature Fusion Module to build a bridge mechanism with both structure awareness and semantic consistency between the encoder and decoder, so as to realize the collaborative modeling of semantic information and edge information. Figure 4, the module jointly models the decoder feature F'_{De}, the encoder semantic feature F'_{En} and the edge feature F'_E through the multi-head attention mechanism, enhances the semantic consistency and boundary perception ability, and realizes the deep interaction between edge and semantic information.

In order to achieve efficient fusion of multiple features in interaction, Point-Wise Convolutio was used to channel map the decoded feature F'_{De}, encoded feature F'_{En} and edge feature F'_E respectively. Depthwise Separable Convolution is used to establish local spatial context and maintain the independence between channels. In the multi-head attention computation stage,

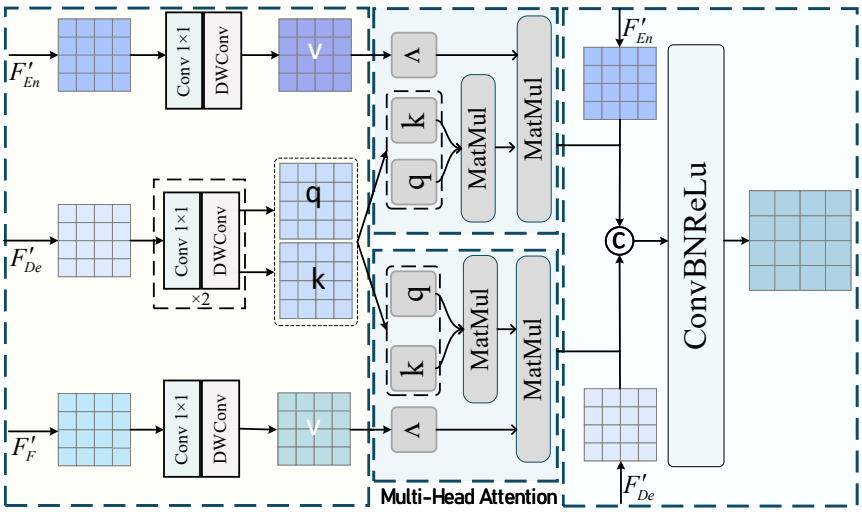

Figure 4. MFFM Module. F'_{De}, F'_{En},F'_E respectively represent the encoding feature, decoding feature, and edge enhancement feature.

the decoder feature F'_{De} generates Q and K representations simultaneously, while the encoder feature F'_{En} and the edge feature F'_E generate V_{sem} and V_{edge}, respectively. The multi-head attention mechanism establishes the correlation between F'_{De} and V_{sem} and V_{edge} features in the global scope, so as to realize the interaction between semantic features and edge information.

Finally, the two attention enhancement results are cascaded with the original input features F'_{De} and F'_{En}, and the multi-source information is uniformly encoded through the convolution fusion module, so that the fusion features have both semantic and edge information.

To achieve this, based on the decoder feature $\mathbf{F}'_{\mathrm{De}} \in \mathbb{R}^{C\times H\times W}$, we use a combination of point-wise convolution and depthwise separable convolution. The joint modeling of query $\mathbf{Q}$ and key $\mathbf{K}$ is formulated as:

$$\begin{aligned}\mathbf{Q} &= \mathrm{DWConv}\big(\mathrm{Conv}_{1\times1}(\mathbf{F}'_{\mathrm{De}})\big)\\ \mathbf{K} &= \mathrm{DWConv}\big(\mathrm{Conv}_{1\times1}(\mathbf{F}'_{\mathrm{De}})\big)\end{aligned} \tag{7}$$

In the Value branch, the module designs a dual guidance strategy for semantic and edge. The semantic-guided branch uses the encoder to output features $F'_{En} \in \mathbb{R}^{C\times H\times W}$, The edge-guided branch makes use of $F'_E \in \mathbb{R}^{C\times H\times W}$, which extracts the semantic context representation through a combination of point-wise convolution and deep convolution:

$$\begin{aligned}V_{sem} &= \mathrm{DWConv}(\mathrm{Conv}_{1\times1}(\mathbf{F}'_{\mathbf{En}}))\\ V_{edge} &= \mathrm{DWConv}(\mathrm{Conv}_{1\times1}(\mathbf{F}'_{\mathbf{E}}))\end{aligned} \tag{8}$$

Then, in order to realize the collaborative fusion of semantic information and edge structure, (Q,K,V_{sem}) and (Q,K,V_{edge}) are respectively applied to calculate the attention weights of semantics and edge, which are formally expressed as:

$$\begin{aligned}O_{sem} &= \mathrm{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V_{sem}\\ O_{edge} &= \mathrm{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V_{edge}\end{aligned} \tag{9}$$

Here d_k denotes the dimension of the key vectors. The enhanced features for the semantic path O_{sem} and the edge path O_{edge} are obtained, respectively.

Subsequently, the two in-context enhanced features are cascaded and fused with the encoder feature and the decoder feature. Further integration is achieved through1×1 convolution and batch normalization layers to form a unified representation that integrates global semantics, local structure, and edge detail information. This fusion process effectively enhances the model's capabilities in multi-scale context awareness and boundary detail capture.

Multi-Level Fusion Segmentation Head

In pixel-level semantic segmentation tasks, shallow features contain rich spatial structure details, while deep features represent abstract high-level semantics, which are naturally complementary to each other. How to effectively fuse these features is crucial to improve the prediction accuracy. Especially in the process of multi-scale feature integration, taking into account both local details and global semantics is helpful to enhance the model's ability to perceive objects of different scales. However, existing methods often rely on the single-scale output of the decoder, which is difficult to fully capture multi-scale semantic information, thus limiting the expressive power of key features.

To this end, a multi-scale feature fusion strategy is adopted: the three different levels of feature maps $F_i$extracted by the decoder are unified to the same resolution by convolution and upsampling operations, element-by-element addition is performed, and the final segmentation result is obtained by convolution:

$$Y = \mathrm{Conv}\left(\sum_{i=1}^{3} \mathrm{Up}(\mathrm{Conv}(F_i))\right), i \in \{1,2,3\} \tag{10}$$

Here Y denotes the segmentation result,and UP represents upsampling.

Loss Function

We adopted a composite Loss function L_{total}, combining the Soft Cross-Entropy Loss with the Dice Loss to balance pixel-level classification accuracy and region-level segmentation performance, thereby achieving a more balanced and stable model training.

$$\mathscr{L}_{total} = \mathscr{L}_{ce} + \mathscr{L}_{dice} \tag{11}$$

Cross-entropy loss is defined as:

$$\mathscr{L}_{ce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^n \log \hat{y}_k^n \tag{12}$$

Here $\mathscr{L}_{ce}$ measures the discrepancy between the predicted class probabilities $\hat{y}_k^n$ and the ground truth labels y_k^n for N samples, K denotes the number of classes.

To address the issue of class imbalance, we introduce the Dice loss:

$$\mathscr{L}_{dice} = -\frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\hat{y}_k^n y_k^n}{\hat{y}_k^n + y_k^n} \tag{13}$$

The Dice loss emphasizes high-confidence predictions, thereby enhancing model performance in the presence of class imbalance. This combined approach optimizes the overlap between predicted and ground truth regions, improving pixel-level classification accuracy and segmentation quality.

EXPERIMENTS

In this section, the experimental setup is detailed, including the datasets employed, experimental details, and evaluation metrics. Subsequently, we design a series of ablation experiments to systematically compare and evaluate the performance of the model, highlight the role of each key module in the overall framework, and verify the effectiveness and advantages of the proposed method.

Datasets

We perform on three diverse and challenging benchmark datasets: ISPRS Vaihingen, ISPRS Potsdam, and Agriculture-Vision[6]. These datasets cover a variety of typical scenes such as cities, towns, and basic farmland, containing multiple land cover types, environmental conditions, and scenes at different levels. Through experimental verification on diverse datasets, we ensure the generality and robustness of the proposed method in a wide range of application scenarios.

ISPRS Vaihingen and Potsdam Datasets

The Vaihingen dataset contains 33 high-resolution TOP images (GSD 9 cm, average size of 2494×2064) with five foreground classes and one background class, of which 16 are used for training and 17 for testing.The Potsdam dataset contains 38 ultra-high resolution TOP images (GSD 5 cm, pixels of 6000×6000) with the same category as Vaihingen, cropped to 1024×1024 and then used, 24 of which are used for training. Fourteen images are used for testing.

Agriculture-Vision Datasets

It is a large-scale agricultural aerial dataset collected from multiple agricultural areas in the United States from 2017 to 2019. This study uses the 2019 section with a total of 22627 (pixels of 512×512) images covering seven categories: Background(BG),Planter Skip(PS), Water(WT), Weed Cluster(WC), Waterway(WW), and Nutrient Deficiency(ND), of which 14628 were used for training, 3779 for validation, and 4220 for testing.

Experimental Setup

Data Preparation: For the Vaihingen and Potsdam datasets, images were cropped to 1024×1024 pixels, while the Agriculture-Vision dataset was resized to 512×512 pixels.

Training Configuration: The AdamW optimizer with a cosine learning rate schedule was employed, using a base learning rate of 6×10^{-4}. The models were trained on RTX A40 GPUs under Ubuntu 20.04. For the Vaihingen and Potsdam datasets, training was conducted for 105 epochs with data augmentation including random flipping, scaling, and cropping for the Agriculture-Vision dataset, training was performed for 50 epochs.

Model Initialization: The backbone network was initialized with pre-trained CSWin weights, while the decoder was randomly initialized.

Comparison Methods: Several state-of-the-art image segmentation methods were selected for comparison, with a pure convolutional UNet serving as the baseline, which models only local context at each stage.

Evaluation Metrics

To evaluate segmentation performance, this study employs Overall Accuracy (OA), F1, and mean Intersection over Union (mIoU), defined as follows:

$$\mathrm{OA} = \frac{TP+TN}{TP+FP+TN+FN} \tag{14}$$

$$\mathrm{F1} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{2TP}{2TP+FP+FN} \tag{15}$$

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{16}$$

Here k is the number of target segmentation categories, TP is the true positive, TN is the true negative, FP is the false negative, and FN is the false negative pixel numbers of the result.

Comparative Experiments

To verify the validity of this model, we compared it with the latest methods on three widely used open access datasets.

Model	Im_Surf	Building	LowVeg	Tree	Car	F1 (%)	mIoU (%)	OA (%)
FCN[15]	91.90	94.53	83.05	89.73	87.56	89.35	80.98	89.94
DeeplabV3+[17]	92.48	94.82	83.88	89.95	87.48	89.72	81.58	90.46
PSPNet[18]	90.35	93.93	83.31	89.05	72.12	85.75	75.80	89.11
FT-Unetformer[42]	93.46	95.95	85.60	90.68	**91.83**	91.17	83.97	91.57
ABCNet[20]	92.82	95.36	85.05	90.58	89.68	90.70	83.17	91.04
DCSwin[43]	93.29	95.91	85.35	90.76	89.69	91.00	83.68	91.45
CMTFNet[25]	93.32	95.87	85.57	90.74	90.53	91.21	84.02	91.48
SFFNet[44]	93.37	95.72	85.16	90.71	90.99	91.19	84.01	91.42
MAResUnet[1]	93.44	95.63	85.85	90.96	90.54	91.29	84.13	91.60
A2FPN[45]	93.32	**96.00**	85.30	90.30	89.75	90.93	83.58	91.43
MANet[46]	93.30	95.44	85.20	90.76	90.44	91.03	83.72	91.27
RS3Mamba[35]	93.04	95.75	85.02	90.76	90.19	90.95	83.60	91.30
MF-Mamba[47]	93.40	95.23	85.34	90.80	91.45	91.40	84.35	91.59
MFEF-UNet(Ours)	**93.90**	95.93	**85.96**	**91.06**	90.93	**91.56**	**84.61**	**91.84**

Table 1. Experimental Results of Different Models on the Vaihingen Dataset

Results on the Vaihingen Dataset: The evaluation results of various methods on the Vaihingen dataset are listed in table 1, indicating that MFEF-UNet outperforms existing comparison methods in all metrics (F1, mIoU, and OA). we show the visualization of segmentation results of different models in Figure 5. In contrast, DeepLabv3+ is more advantageous in edge detail capture by using its atrous spatial Pyramid Pooling. A2-FPN introduces an attention-enhanced feature pyramid to

effectively fuse multi-scale context information, while MAResUNet and MANet strengthen the collaborative modeling of local and global semantics through different attention mechanisms. SFFNet uses frequency domain features to improve the accuracy of boundary segmentation, and its performance is stable. However, ABCNet and DCSwin achieve a good balance by designing lightweight attention structures. In terms of the emerging Transformer architecture, FT-Unetformer adopts Swin Transformer as the feature extraction encoder and combines the global-local attention modeling mechanism to achieve a good balanced performance in multiple categories. CMTFNet strengthens the cross-channel information interaction mechanism and improves the context expression ability. In MFEF-UNet, we introduce edge branch and semantic fusion mechanism, which effectively improves the modeling ability of the model for fine-grained structures. We visualize the predictions of some models in typical scenes, such as high-density building areas and small object distribution areas, and highlight regions of interest in purple for visibility and contrast. It can be seen from the figure that the MFEF-UNet model exhibits higher segmentation accuracy and completeness in building boundaries, complex contour structures (such as long boundary walls), and dense small objects (such as vehicles) recognition tasks. Especially in areas with low contrast or complex backgrounds, MFEF-UNet model can focus on key structures more accurately, showing stronger detail modeling ability, edge analysis ability, spatial perception and attention focusing ability.

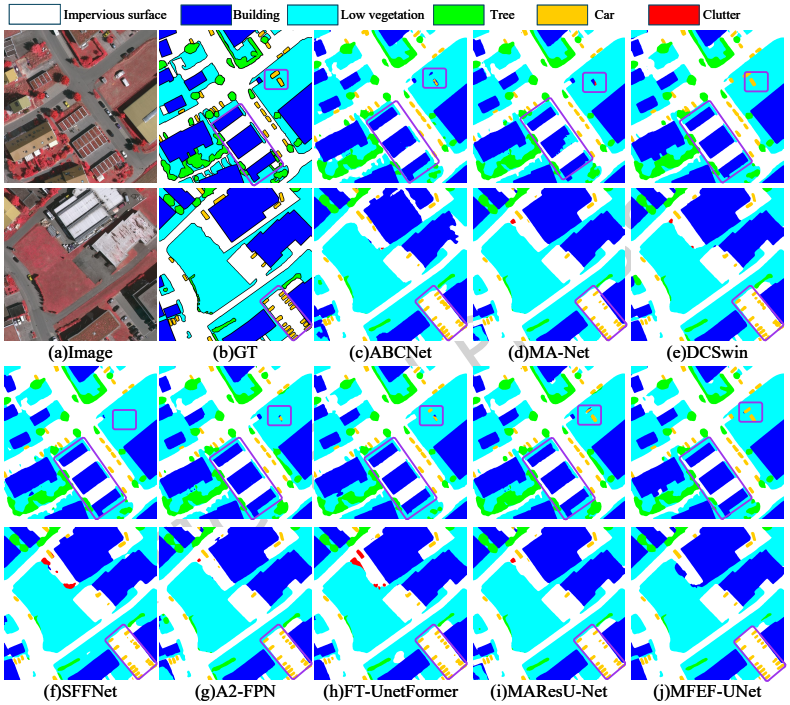

Figure 5. Visualization of Segmentation Results of Different Models on the Vaihingen Dataset

Results on the Postdam Dataset: To evaluate the generalization performance of the proposed method under different scenes and spatial resolutions, we conducted additional experiments on the Potsdam dataset, and the results are shown in Table 2. The MFEF-UNet model achieves an average F1 of 93.10%, mIoU of 87.30%, and OA of 91.74% on this dataset, which exceeds all comparison methods. Different from the Vaihingen dataset, Potsdam provides more abundant training samples, which can more comprehensively reflect the change characteristics of different land cover types.

To evaluate the generalization performance of the proposed method under different scenes and spatial resolutions. we conducted additional experiments on the Potsdam dataset, and the results are shown in Table 2. The MFEF-UNet model achieves an average F1 of 93.10%, mIoU of 87.30%, and OA of 91.74% on this dataset, which exceeds all comparison methods. Different from the Vaihingen dataset, Potsdam provides more abundant training samples, which can more comprehensively reflect the change characteristics of different land cover types. In order to further verify the segmentation performance of each model in complex scenes, we show the visualization of segmentation results of different models in Figure 6. Areas of interest, including complex buildings and low-rise vegetation areas, are highlighted in purple in the visualization to highlight differences in the performance of each method in detail areas. Although the Potsdam dataset provides rich high-resolution remote sensing samples. MAResUNet, MANet perform relatively stable in enhancing local attention expression and recovering the main structure in large-scale object segmentation. However, when dealing with building occlusion and complex contour

Model	Im_Surf	Building	LowVeg	Tree	Car	F1(%)	mIoU(%)	OA (%)
FCN[15]	90.59	92.84	85.33	86.25	92.80	89.64	81.37	88.34
DeeplabV3+[17]	92.85	95.41	86.94	87.67	95.24	91.62	84.75	90.27
PSPNet[18]	91.32	95.20	85.46	87.46	86.15	89.12	80.58	89.25
FT-Unetformer[42]	94.34	**97.12**	88.17	88.95	96.39	92.99	87.13	91.66
ABCNet[20]	93.39	95.98	87.09	88.06	94.83	91.87	85.17	90.58
DCSwin[43]	94.43	96.89	87.71	**89.26**	95.58	92.78	86.74	91.65
CMTFNet[25]	94.00	96.58	88.02	89.02	**96.56**	92.83	86.85	91.45
SFFNet[44]	**94.50**	97.01	87.87	88.25	96.66	92.86	86.93	91.59
MAResUnet[1]	93.60	96.34	87.48	87.93	95.82	92.24	85.82	90.93
A2FPN[45]	93.87	96.12	87.57	88.22	95.63	92.28	85.89	91.01
MANet[46]	93.82	96.10	87.93	88.78	95.59	92.44	86.14	91.20
RS3Mamba[35]	93.41	95.93	87.31	88.35	95.91	92.18	85.71	90.75
MF-Mamba[47]	93.33	96.78	87.73	88.45	95.88	92.44	86.16	91.17
MFEF-UNet(Ours)	94.44	96.91	**88.52**	89.25	96.08	**93.10**	**87.30**	**91.74**

Table 2. Experimental Results of Different Models on the Postdam Dataset.

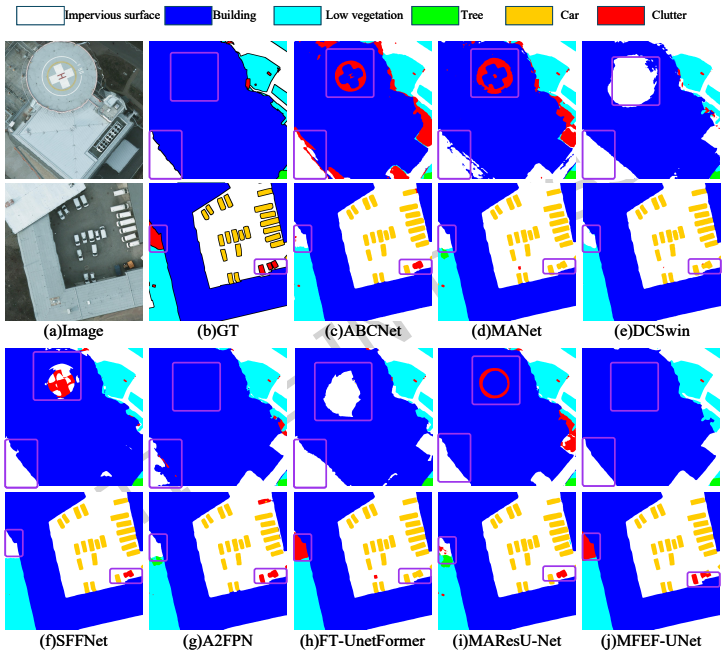

Figure 6. Visualization of Segmentation Results of Different Models on the Potsdam Vision Dataset

structures,MAResUnet, MANET perform relatively stable in enhancing local attention expression. There are still semantic breaks, which lead to recognition errors. A2FPN is more balanced in the overall structure modeling, and can effectively integrate semantic information of different scales to achieve good global perception ability. However, it is still insufficient in the modeling ability of fine structure. Structures such as CMTFNet and FT-Unetformer strengthen the global feature modeling ability by introducing a Transformer encoder, and perform well in the segmentation of complex semantic regions. However, there is still information attenuation in the edge transition region, which affects the detail retention ability. SFFNet improves edge perception by introducing frequency domain features, which is suitable for processing structured objects. However, SFFNet relies on spectrum representation and is prone to confusion segmentation in complex texture regions. In contrast, MFEF-UNet effectively enhances the ability of local detail modeling by introducing the sub-block segmentation strategy, enabling the model to maintain global consistency while taking into account the fine analysis of local structures. Thus, it demonstrates stronger spatial perception and semantic consistency. In the tasks of complete building segmentation and small target edge recognition, the model has achieved more complete and smooth segmentation effects, and its performance is superior to all existing comparison methods.

Results on the Vaihingen Agriculture-Vision Dataset: To further verify the generalization ability of the model, we introduce

Model	BG	PS	WT	WC	WW	ND	F1 (%)	mIoU (%)	OA (%)
FCN[15]	93.97	81.31	89.28	78.58	81.93	77.24	83.72	72.47	90.80
DeeplabV3+[17]	92.65	75.85	88.01	75.48	78.49	72.78	80.54	68.07	88.89
PSPNet[18]	92.59	62.70	87.39	74.67	78.20	74.11	81.23	65.33	89.45
FT-Unetformer[1]	94.72	85.75	91.56	82.32	**85.51**	80.74	86.77	76.97	92.05
ABCNet[20]	90.39	56.70	76.90	66.42	67.46	65.86	70.62	55.71	85.14
DCSwin[43]	94.56	84.57	91.30	81.90	84.38	79.90	86.10	75.97	91.80
CMTFNet[25]	93.35	80.07	89.03	77.09	79.53	75.24	82.39	70.60	89.93
SFFNet[44]	94.80	85.81	91.26	82.70	85.06	80.62	86.71	76.87	92.14
MARseUnet[42]	92.56	72.89	87.51	74.91	78.11	74.58	80.11	67.47	88.81
A2FPN[45]	92.95	77.20	87.65	76.31	76.10	74.25	80.74	68.31	89.30
MANet[46]	92.59	72.52	86.82	74.16	74.43	73.34	78.98	65.99	88.72
RS3Mamba[35]	92.60	76.35	87.43	75.11	78.44	75.32	80.88	68.46	88.92
MF-Mamba[47]	93.81	79.35	89.92	78.21	81.20	77.29	83.23	71.89	90.59
MFEF-UNet (Ours)	**94.86**	**86.18**	**91.61**	**82.89**	85.47	**81.06**	**87.01**	**77.39**	**92.25**

Table 3. Experimental Results of Different Models on the Agriculture Vision dataset.

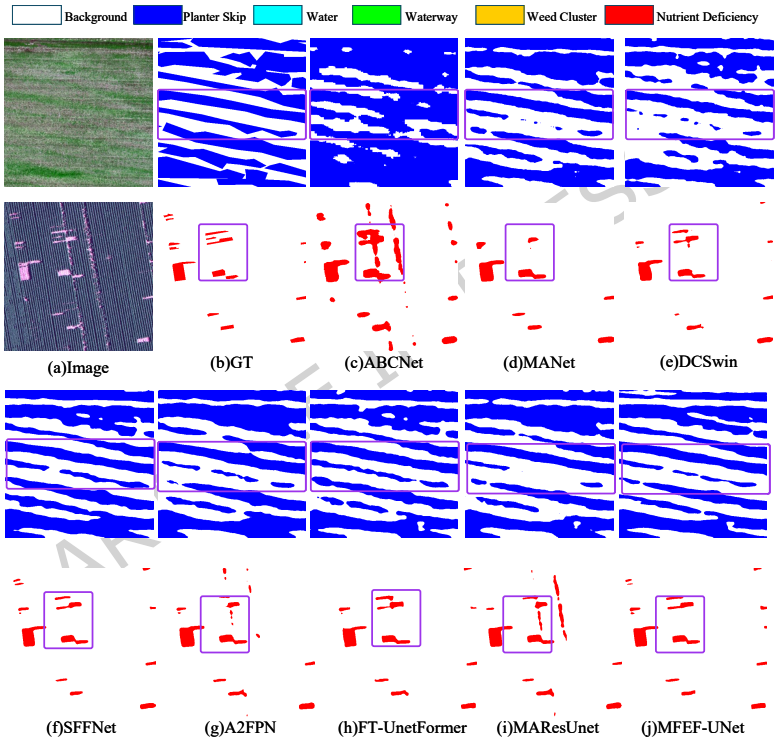

Figure 7. Visualization of Segmentation Results of Different Models on the Agriculture Vision Dataset

independent farmland areas, which are significantly different from typical urban scenarios such as Vaihingen and Potsdam, as test scenarios to evaluate the performance of the model in unstructured environments. Table 3 MFEF-UNet model still shows superior performance in this type of agricultural scenario, with mIoU of 77.39%, F1 of 87.01%, and overall OA of 92.25%. In terms of class balance and robustness, the model maintains high stability. In particular, on the long and narrow structures common in agricultural scenarios, the MFEF-UNet model significantly outperforms most of the comparison methods in terms of boundary continuity and region identification accuracy.

To gain a deeper understanding of the segmentation ability of each model in different scenarios, we show a visualization of the segmentation results of multiple models in Figure 7, highlighted in purple for comparison. From the visualization results, it can be observed that different methods have significant differences in edge preservation and detail recognition. In the advantage of having a large amount of training data, ABCNet shows obvious disadvantages. Especially, the segmentation effect is poor in the regions with significant local changes such as nutrient deficiency, which verifies the lack of modeling ability

in high-resolution remote sensing images. MANet fails to capture complex structures well, and the segmentation edges are blurred and broken. A2-FPN achieves relatively balanced segmentation results on multiple categories, but its resolution is limited in processing high-detail regions, and it is difficult to accurately recover the boundary structure. The models based on spatial features, such as MANet and CMTFNet, have certain ability in maintaining structural integrity, but there are still errors in semantic consistency and boundary continuity.

In contrast, our proposed model MFEM-UNet combined with cross-scale context fusion module and edge enhancement mechanism can effectively capture structural information while maintaining global semantic consistency.

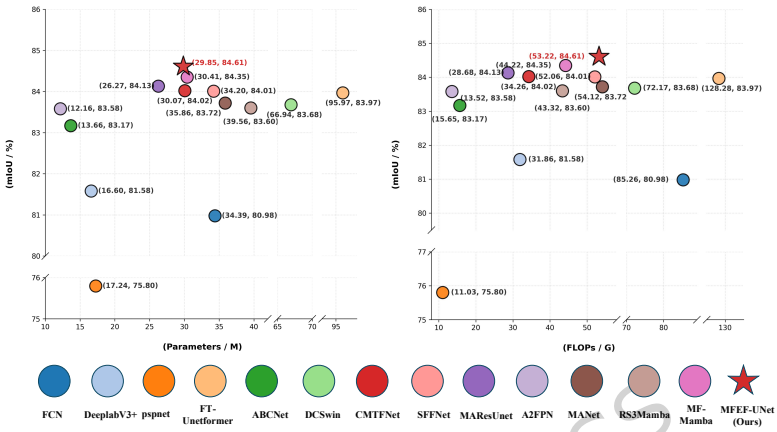

Figure 8. Visualization of comparison with state-of-the-art network parameter count and floating-point arithmetic on the vaihingen dataset

To quantitatively evaluate the computational efficiency of different methods, the Figure 8 shows the number of parameters (Params) and floating-point operations (FLOPs) of each model. It can be seen that traditional convolutional networks (such as FCN and DeepLabV3+) have low overall computational overhead, while transformer-based methods (such as FT-Unetformer) introduce significant computational cost while improving modeling ability, with FLOPs as high as 128.28G. In contrast, lightweight models such as ABCNet and A2FPN show obvious advantages in parameter scale and computational complexity, but their feature representation ability is relatively limited.

In the Mamba family of methods, RS3Mamba has 39.56M parameters and 43.32G computation, respectively, while the proposed method only needs 29.85M parameters and achieves good results while maintaining similar computational complexity (53.22G FLOPs). In general, the proposed method achieves a relatively balanced result between model parameters and computational efficiency.

Ablation Experiments

To evaluate the effectiveness of the individual components in MFEF-UNet, we conducted systematic ablation experiments on the ISPRS Vaihingen dataset. The evaluation focuses on four key performance metrics: mIoU and F1, Params, FLOPS. In Table 4, A is the baseline model, (B-J) is the combination of modules, and new components are represented by ✓. All the results are the average of multiple independent runs to ensure the robustness and reliability of the experimental conclusions.

Composition analysis of MFEF-UNet: To verify the independent contributions of each module, key modules were gradually added to the baseline model constructed after all other components were removed for evaluation. This baseline only retains the CSWin backbone and the CNN-based decoder. The contribution results of each module are summarized in the Table 4. Figure 9 highlights the impact of adding different module combinations of MFEF-UNet to the baseline model, demonstrating the effectiveness of the method in context feature extraction and edge detail enhancement.

1) The influence of the LG-FEM module: Experimental results show that after adding LG-FEM to the baseline model, mIoU increased to 83.61% and F1 increased to 90.96%, with improvements in both indicators. In Figure 9(k), compared with the baseline, the model's recognition of small targets (such as cars) is more accurate, and the boundary clarity is enhanced.

2) Influence of MFFM and EEM modules: The MFFM aims to improve the representation ability of fine-grained structures and edge objects. By guiding efficient interaction of multi-scale features. When the baseline model only introduces MFFM, the mIoU of the model is increased to 83.43%, and the F1 is increased to 90.85%. As shown in Figure 9(i), for the baseline model, introducing only the MFFM module leads to a certain misjudgment in small target recognition. After the introduction of EEM, the mIoU of the model is increased to 83.68%, and F1 is increased to 91.01%, as shown in Figure 9(j). This kind of

Model	LG-VSSM	MFFM	EEM	MLF-Head	mIoU	F1	Params (M)	FLOPS (G)
A					82.55	90.31	**23.69**	**30.70**
B	✓				83.61	90.96	25.04	33.04
C		✓	✓		83.68	91.01	28.39	41.40
D		✓			83.43	90.85	28.70	49.32
E				✓	83.32	90.79	24.16	39.41
F	✓	✓	✓		83.26	91.24	29.71	45.30
G	✓			✓	83.96	91.17	25.50	42.52
H		✓	✓	✓	84.04	91.22	28.40	49.33
I	✓	✓		✓	84.07	91.24	26.89	45.30
J	✓	✓	✓	✓	**84.61**	**91.56**	29.85	53.22

Table 4. Impact of different module combinations on model performance and computational cost.

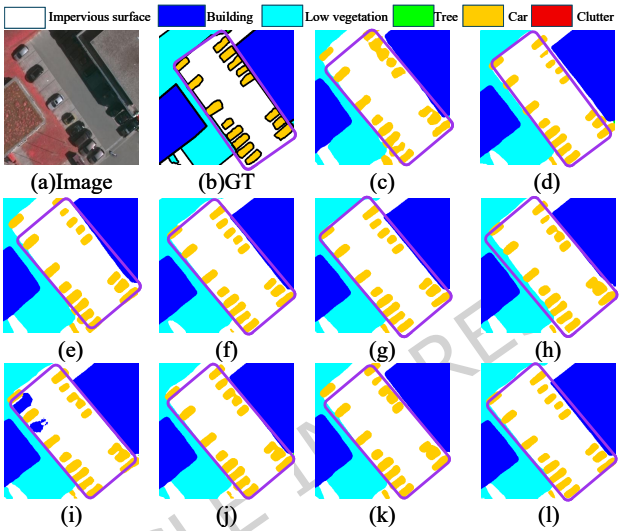

Figure 9. Visualization of the segmentation performance with different module combinations on the ISPRS Vaihingen dataset, focusing on the enlarged local regions. (a) Original image. (b) Ground Truth. (c) Baseline. (d) w/o MLF-Head. (e) w/o MFFM. (f) w/o EEM. (g) w/o LG-FEM. (h) MLF-Head branch. (i) MFFM branch. (j) MFFM+EEM branch. (k) LG-VSSM branch. (l) MFEF-UNet.

misjudgment is significantly reduced, and the segmentation effect of edge structure and fine-grained objects is significantly improved. The experimental results show that the MFFM module is effective in detail enhancement and small target recognition after merging the edge features of EEM.

3) Impact of MLF-Head module: MLF-Head is a segmentation head module that fuses multi-layer decoder feature maps, aiming to refine boundary representation and improve semantic consistency. After the introduction of this module, mIoU is increased to 83.32% and F1 is increased to 90.79%, which forms an effective supplement to the overall performance in the feature decoding stage. Further combined with LG-FEM and MFFM+EEM, the model achieves the best performance mIoU 84.61%, F1 91.56 %, which fully proves its irreplaceable in the final feature fusion and boundary optimization.

To verify the feature chunking strategy, we set different number of chunking to explore the best trade-off between model computational overhead and performance. Table 5 shows the experimental results with different number of blocks. The results show that the introduction of blocking mechanism can improve the performance of the model, and all evaluation indicators are better than the baseline model without blocking, which further verifies the effectiveness of the strategy.

When α is set to 4, the model achieves the best results, with the highest mIoU and F1 values. Compared with no block or less block $\alpha = 1$, $\alpha = 2$, $\alpha = 4$ improves the segmentation accuracy with only a reasonable amount of calculation. Compared with $\alpha = 6$, $\alpha = 4$ not only keeps the performance advantage, but also effectively avoids the extra computation cost. This result shows that a reasonable blocking strategy can achieve the best balance between model performance and computational efficiency.

The LG-FEM module is designed based on the VSSM architecture and aims to further improve the segmentation performance

α	mIoU (%)	Params (M)	FLOPS (G)
1	84.35	29.85	**52.52**
2	84.46	29.85	53.13
4	**84.61**	29.85	53.22
6	84.45	29.85	53.40

Table 5. The influence of different sub-block quantities in LG-FEM on model performance

Method	mIoU	Params (M)	FLOPS (G)
VSSM[33]	84.31	**29.73**	**51.26**
LG-VSSM	84.49	29.74	51.97
LG-FEM	**84.61**	29.85	53.22

Table 6. Compare to the basic VSSM module

through the chunking mechanism and feature enhancement. To verify the effectiveness of the proposed module with respect to the standard VSSM, we performed ablation experiments on the LG-FEM. Table 6 shows the comparison of segmentation performance under different Settings, including baseline VSSM, LG-VSSM, and LG-FEM. The experimental results show that compared with the standard VSSM module, the introduction of the blocking mechanism LG-VSSM can improve the mIoU of the model, while the amount of calculation increases slightly. However, after adding feature enhancement and CBAM, the model achieves a more significant improvement in mIoU. Compared with the standard VSSM, the mIou of the two algorithms are increased by about 0.18 and 0.30 percentage points respectively, and the increase of Parameter number and Flops is small. It provides a reliable basis for further improving the VSSM architecture.

Method	mIoU	F1	0A	Params (M)	FLOPS (G)
EESM	84.32	91.39	91.69	29.85	53.63
EEM	84.61	91.56	91.84	29.85	53.22

Table 7. Comparison of EESM and EEM modules

To verify the effectiveness of the Sobel and Laplacian edge operators, we designed a substitution experiment. We build a variant of the model EESM in which the original Sobel and Laplacian modules are replaced by a Standard Convolution Block (SCB). This SCB consists of 3×3 depthwise separable convolutions, BN and ReLU activations. The experimental results are shown in the table 7 .

Although standard convolutional layers are able to implicitly learn edge features, they usually struggle to distinguish between high-frequency structural boundaries and complex texture noise in remote sensing images. EEM introduces a strong inductive bias by combining fixed operators, the Sobel operator and the Laplacian operator, with learnable scaling factors. This allows the network to explicitly focus its attention on gradient information, thus ensuring that boundary features are preserved and prioritized rather than being obscured by rich semantic features learned at deeper layers. This design effectively acts as a "soft" prior, combining the reliability of classical edge detection with the adaptability of deep learning.

CONCLUSION

In this study, we designed the MFEF-UNet semantic segmentation method for high-resolution remote sensing images based on the structure of the encoder and decoder. In the MFEF-UNet model, the sub-block segmentation strategy is used in LG-FEM to enhance the recovery ability of local details of the model. The efficient fusion of multi-scale features by MFFM enhances the recognition ability of fine-grained targets. At the same time, EEM is introduced to improve the accuracy of segmentation boundaries by explicitly enhancing the edge response. The above design effectively promotes the complementary interaction between global and local features, as well as semantic and edge information. A large number of experiments on three key datasets show that the proposed MFEF-UNet outperforms the existing state-of-the-art methods in both segmentation accuracy and generalization ability. We admit that the scale fusion and edge enhancement modules introduced in the MFEF-UNet bring certain computational and storage overhead within an acceptable range. In the future, we will explore more lightweight network designs to enhance real-time processing performance, expand multi-modal remote sensing data, and further leverage its potential in multi-source data fusion.

Data Availability

The datasets analyzed during this study are available in the following public domains: https://github.com/SHI-Labs/Agriculture-Vision and https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx.

Funding

This work is a research achievement supported by the National Key R&D Program of China Major Project (No. 2022ZD0115800) and the National Natural Science Foundation of China (No. 62262065).

References

1. Li, R. *et al.* Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geosci. Remote. Sens.* **60**, 1–13 (2021).
2. Yuan, X., Shi, J. & Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert. Syst. with Appl.* **169**, 114417 (2021).
3. Pan, L. *et al.* M 3-cr: Multi-scale multi-branch mamba for sar-assisted optical image thick cloud removal. *IEEE Transactions on Geosci. Remote. Sens.* (2025).
4. Dong, R. *et al.* High-resolution land cover mapping through learning with noise correction. *IEEE Transactions on Geosci. Remote. Sens.* **60**, 1–13 (2021).
5. Li, Z., Zhu, Q., Yang, J., Lv, J. & Guan, Q. A cross-domain object-semantic matching framework for imbalanced high spatial resolution imagery water-body extraction. *IEEE Transactions on Geosci. Remote. Sens.* **62**, 1–15 (2024).
6. Chiu, M. T. *et al.* Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2828–2838 (2020).
7. He, S. *et al.* A distinctive eocene asian monsoon and modern biodiversity resulted from the rise of eastern tibet. *Sci. Bull.* **67**, 2245–2258 (2022).
8. Zhang, X., Yuan, G., Hua, Z. & Li, J. Tsmga: temporal-spatial multi-scale graph attention network for remote sensing change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* (2025).
9. Ji, H., Xie, F., Pan, L., Zheng, Y. & Shi, Z. Huntnet: Homomorphic unified nexus topology for camouflaged object detection. *IEEE Transactions on Image Process.* (2025).
10. Ma, X. *et al.* Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Transactions on Geosci. Remote. Sens.* (2024).
11. Zhang, X., Dong, K., Cheng, D., Hua, Z. & Li, J. Stwanet: Spatio-temporal wavelet attention aggregation network for remote sensing change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* (2025).
12. Pan, L., Zhang, X., Xie, F., Zhang, H. & Zheng, Y. Sgiqa: semantic-guided no-reference image quality assessment. *IEEE Transactions on Broadcast.* (2024).
13. Liu, X., Jiao, L., Li, L., Tang, X. & Guo, Y. Deep multi-level fusion network for multi-source image pixel-wise classification. *Knowledge-Based Syst.* **221**, 106921 (2021).
14. Wu, H., Zhang, M., Huang, P. & Tang, W. Cmlformer: Cnn and multiscale local-context transformer network for remote sensing images semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **17**, 7233–7241 (2024).
15. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440 (2015).
16. Gao, Y. *et al.* Semantic segmentation of remote sensing images based on multiscale features and global information modeling. *Expert. Syst. with Appl.* **249**, 123616 (2024).
17. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis machine intelligence* **40**, 834–848 (2017).
18. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890 (2017).

19. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).

20. Li, R. *et al.* Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS journal photogrammetry remote sensing* **181**, 84–98 (2021).

21. Yu, B., Yang, L. & Chen, F. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **11**, 3252–3261 (2018).

22. Zhou, Y., Xia, H., Yu, D., Cheng, J. & Li, J. Outlier detection method based on high-density iteration. *Inf. Sci.* **662**, 120286 (2024).

23. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).

24. He, X. *et al.* Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE transactions on geoscience remote sensing* **60**, 1–15 (2022).

25. Wu, H., Huang, P., Zhang, M., Tang, W. & Yu, X. Cmtfnet: Cnn and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Transactions on Geosci. Remote. Sens.* **61**, 1–12 (2023).

26. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

27. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).

28. Xie, E. *et al.* Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. neural information processing systems* **34**, 12077–12090 (2021).

29. Strudel, R., Garcia, R., Laptev, I. & Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272 (2021).

30. Chen, J. *et al.* Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).

31. Zhang, R., Zhang, Q. & Zhang, G. Lsrformer: Efficient transformer supply convolutional neural networks with global information for aerial image segmentation. *IEEE Transactions on Geosci. Remote. Sens.* **62**, 1–13 (2024).

32. Zhang, X., Wang, Z., Li, J. & Hua, Z. Mvafg: Multiview fusion and advanced feature guidance change detection network for remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **17**, 11050–11068 (2024).

33. Gu, A. & Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

34. Liu, M. *et al.* Cm-unet: Hybrid cnn-mamba unet for remote sensing image semantic segmentation. *arXiv preprint arXiv:2405.10530* (2024).

35. Ma, X., Zhang, X. & Pun, M.-O. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote. Sens. Lett.* **21**, 1–5 (2024).

36. Wang, Z., Zheng, J.-Q., Zhang, Y., Cui, G. & Li, L. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079* (2024).

37. Liu, J. *et al.* Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, 615–625 (Springer, 2024).

38. Li, L., Yi, J., Fan, H. & Lin, H. A lightweight semantic segmentation network based on self-attention mechanism and state space model for efficient urban scene segmentation. *IEEE Transactions on Geosci. Remote. Sens.* (2025).

39. Chen, K. *et al.* Rsmamba: Remote sensing image classification with state space model. *IEEE Geosci. Remote. Sens. Lett.* **21**, 1–5 (2024).

40. Dong, X. *et al.* Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12124–12134 (2022).

41. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).

42. Wang, L. *et al.* Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote. Sens.* **190**, 196–214 (2022).

43. Wang, L. *et al.* A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* **19**, 1–5 (2022).

44. Yang, Y., Yuan, G. & Li, J. Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *IEEE Transactions on Geosci. Remote. Sens.* (2024).

45. Li, R., Wang, L., Zhang, C., Duan, C. & Zheng, S. A2-fpn for semantic segmentation of fine-resolution remotely sensed images. *Int. journal remote sensing* **43**, 1131–1155 (2022).

46. Li, R., Zheng, S., Duan, C., Su, J. & Zhang, C. Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* **19**, 1–5 (2021).

47. Xiao, P. *et al.* Mf-mamba: Multi-scale convolution and mamba fusion model for semantic segmentation of remote sensing imagery. *IEEE Transactions on Geosci. Remote. Sens.* (2025).

Author contributions statement

Zhang is responsible for manuscript drafting, review, editing, as well as model design and implementation. Yang is responsible for funding acquisition and supervision. Yin is responsible for manuscript review and editing. Chen is responsible for formal analysis and data curation. Wang, Zhao, are responsible for software design. All authors reviewed the manuscript.

Declarations

Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.