Neurocomputing 653 (2025) 131233

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A multimodal fusion framework for semantic segmentation of remote sensing based on multilevel feature fusion learning

Pingzhu Liu

Graduate School at Shenzhen, Tsinghua University, China

ARTICLE INFO

Keywords:
Remote Sensing
Semantic Segmentation
Multimodal Fusion
High-Resolution Image

ABSTRACT

Significant advancements in remote sensing semantic segmentation using many image modalities have been accomplished since the multimodal fusion technique was developed. Nevertheless, existing methods have struggled to consider information from small objects effectively and have been prone to introducing feature noise into the fusion process due to the high number of small objects and the complex architectural textures and color range present in Very High-Resolution (VHR) images. To address these challenges, we have proposed an end-to-end Multi-Modal Fusion Network (MMFNet), which is based on a three-branch fusion framework for leveraging VHR remote sensing images and Digital Surface Model (DSM) data. Recognizing that remote sensing data exhibit a wide variety of colors and intricate textures and that different modalities convey different color and texture information, we have proposed a Local-Global Feature Fusion (LGFF) module to integrate information from different modalities. We have tackled the difficulty of segmenting easily overlooked small objects with limited pixel coverage by devising an Multi-Level Feature Learning (MLFL) module to capture different-level feature representations. Ultimately, the single-modal and multimodal segmentation information are fused in MMFNet to determine the prediction information jointly. Our proposed model attains excellent miou scores of 84.12 % and 86.13 % on the Vaihingen and Potsdam datasets, respectively, effectively establishing a new state-of-the-art benchmark in multimodal image segmentation tasks. These results underscore not only the model's superior performance but also its significant competitiveness in this domain, marking a notable advancement in the field.

1. Introduction

The precise extraction and recognition of man-made architectural elements in urban settings from Very High Resolution (VHR) imagery [1] captured by aerial or satellite platforms is essential for a multitude of applications, including the delineation of road networks [2], the segmentation of buildings [3], the categorization of point clouds [4], the evaluation of image quality [5], the demarcation of water bodies [6], and the semantic analysis of smoke plumes [7]. With the development of Convolutional Neural Networks (CNN), high-resolution remote sensing segmentation has gained tremendous progress. Simultaneously, the integration of Transformer [8] into the realm of remote sensing imagery analysis has garnered considerable interest, representing a new frontier in remote sensing vision capabilities. Despite their excellent performance, these methods consider only local features or introduce significant computational complexity. Moreover, by considering only single-modal image features without multi-modal image features, these methods struggle to achieve high-accuracy remote sensing semantic segmentation. Moreover, these methods only consider high-level image features while ignoring medium- and low-level image features, which is detrimental for small object recognition.

The advancement of computer vision has led to significant progress in semantic segmentation, inspiring developments in remote sensing semantic segmentation. Initially, some researchers utilized deep CNNs [9] to extract image features for image processing tasks. However, CNNs have limitations in considering global features and are prone to overfitting. To address this problem, Vision Transformer (ViT) [10] directly processes images as one-dimensional sequences (patch sequences) and then uses a self-attentive mechanism to process these sequences to capture long-range dependencies and perform image understanding and analysis. Despite its notable achievements, ViT's computational demands and reliance on extensive datasets make it less feasible for deployment across all visual recognition tasks. Consequently, the integration of frameworks like Mamba [11] into the discourse has already yielded promising outcomes, potentially igniting innovative approaches and fresh insights within the research community. In the rapidly developing domain of computer vision, numerous approaches have made significant strides in the area of remote sensing semantic

E-mail address: kevin_ailover@163.com.

https://doi.org/10.1016/j.neucom.2025.131233
Received 5 May 2025; Received in revised form 3 August 2025; Accepted 6 August 2025
Available online 10 August 2025
0925-2312/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

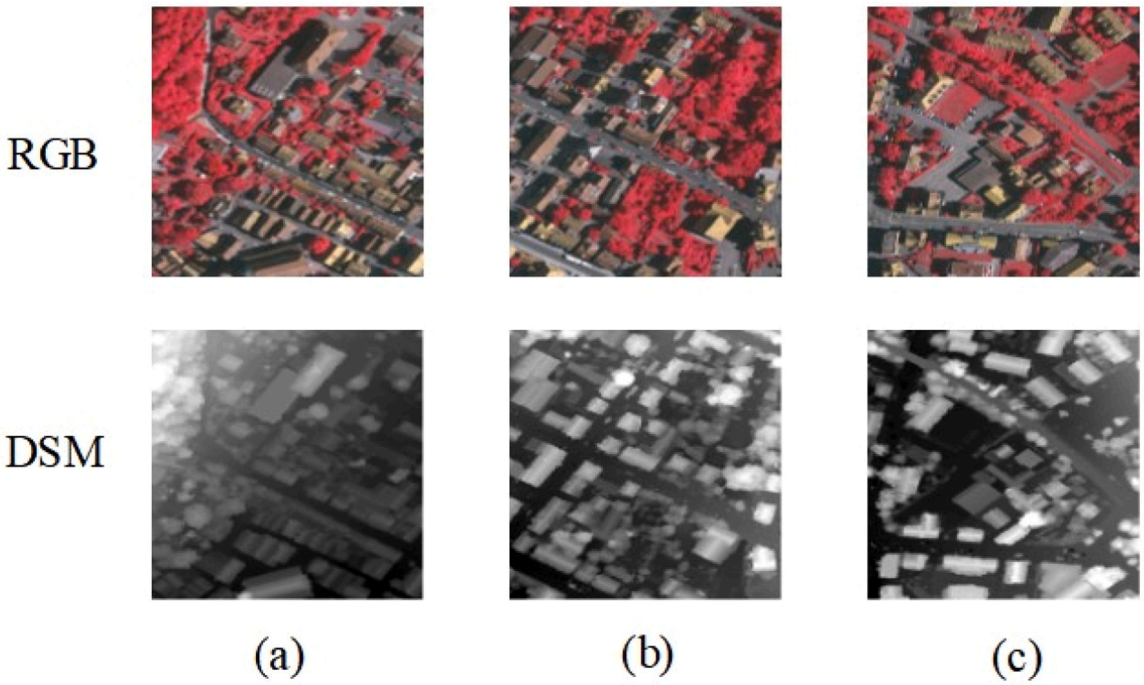

Fig. 1. Examples highlighting the characteristics of remote sensing images, drawn from the ISPRS Vaihingen dataset, illustrate complex architectural textures and small objects.

segmentation. Certain techniques grounded in Convolutional Neural Network (CNN) strategies [12] [13] [14] have demonstrated impressive outcomes. For examples, SegNet [12] proposes a novel convolutional neural network architecture for semantic segmentation. Moreover, U-Net [13] introduces an end-to-end network design that can be optimized during training to yield exceptional performance. Moreover, ST-Unet [14] introduces an innovative semantic segmentation framework for remote sensing images, integrating the Swin Transformer into the traditional CNN-based U-shaped architecture (UNet). Nonetheless, limited research [15] [16] [17] has specifically addressed multimodal fusion, focusing on integrating diverse modalities' unique characteristics.

Multimodal inputs offer a more thorough depiction of the target than single-modal data, which is constrained by the limitations of individual sensors. For instance, in order to investigate the similarities and contrasts between features from the two input modalities for efficient fusion, CIMFNet [15] suggests a multimodal fusion module. Furthermore, an end-to-end cross-modal gated fusion network called CMGFNet [16] is proposed with the goal of obtaining building footprints using digital surface model (DSM) data and very high resolution (VHR) remote sensing photos. Additionally, FTransUNet [17] introduces a multimodal fusion strategy that integrates CNN and Vision Transformer (ViT) across multiple levels within a unified framework for robust semantic segmentation. Although existing methods are already quite good in terms of performance, they generally fail to pay sufficient attention to the importance of learning different feature levels, which is especially critical for improving segmentation accuracy for small objects. Specifically, effectively capturing multi-level features ranging from low-level visual cues to high-level semantic information can significantly enhance a model's ability to understand complex scenes. In addition, existing techniques usually ignore the interactions between different modal data, such as the synergistic effect between RGB and DSM data. Such cross-modal interactions can provide richer and complementary information, which can further enhance the quality and robustness of segmentation results. Taking these factors into account, it is particularly important to develop new strategies to integrate multi-level feature learning and facilitate inter-modal interactions. A selection of data samples from the two datasets are displayed in Fig. 1, illustrating the complex textures typical of remotely sensed images. These images demonstrate the diverse challenges of object segmentation in high-resolution images. The top half of the image presents the object performance in RGB mode, while the bottom half shows the object segmentation in DSM. By comparing the object performance in these two modes, it is clear that our model has the required accuracy and capability to handle complex segmentation tasks. In particular, the need for segmentation of such objects highlights the complexity requirements as well as the efficiency features of our method when confronted with multimodal high-resolution images. This comparison not only emphasizes the importance of achieving high-precision segmentation for efficiently identifying and classifying a wide range of objects, but also demonstrates the effectiveness and reliability of our solution in handling these complex segmentation tasks. It further demonstrates that in the field of multimodal data analysis, our approach is able to provide accurate and reliable segmentation results that meet the high standards required for professional applications.

To address above challenges, we introduce an end-to-end Multi-Modal Fusion Network featuring a three-branch architecture designed to learn multi-level features from very high resolution (VHR) remote sensing images and digital surface model (DSM) data. Specifically, we devise a local-global feature fusion (LGFF) module with both fuse different modalities (e.g., RGB and DSM data) and focus on local-global feature information (e.g., colors and building textures) to improve objects of feature representation and various scales from different stages. Additionally, segmenting small objects is challenging with fewer pixel points resulting in potential information loss during the convolutional extraction process. Consequently, we design a multi-level feature learning (MLFL) module that combines specific low-level features and high-level semantic knowledge to generate an all-encompassing feature representation. Compared with previous three-stage feature fusion networks, such as Feature Pyramid Network (FPN) and Bidirectional Feature Pyramid Network (BiFPN), our designed module realizes four-stage feature fusion for the first time. This innovation not only extends the depth of feature fusion, but also further enhances the feature retention capability and improves the overall model performance

through the effective integration and transmission of multi-level feature information. This multi-level feature fusion approach ensures that information at different levels of abstraction can be fully utilized, thereby improving the model's ability to understand and represent input data with greater accuracy. Our four-stage fusion module aims to solve the problem of information loss that may exist in traditional three-stage architectures and provides a new solution for accurate feature extraction in complex scenarios. Ultimately, we integrate the outputs of single and multi-modal segmentation to collaboratively refine and determine the final prediction. Notably, this work represents our pioneering attempt at implementing a three branch configuration as the terminal output of our model. Comprehensive experiments conducted on the ISPRS Vaihingen and Potsdam datasets demonstrate the superior performance of our MMFNet model compared to existing state-of-the-art (SOTA) methods.

In brief, the following are our main contributions:

- We introduce a new three-branch architectural framework designed for remote sensing semantic segmentation, leveraging multi-modal fusion to combine features extracted from both RGB and DSM imagery.
- We have developed an innovative local-global feature fusion (LGFF) module that combines various stages of RGB and DSM data, incorporating both fine-grained low-level features and contextual high-level features. Furthermore, we have created a new multi-level feature learning (MLFL) module to produce a multi-scale feature representation.
- We perform comprehensive tests on two datasets, proving the superior performance of our network (MMFNet) in semantic segmentation of remote sensing images and outperforming popular state-of-the-art approaches on all assessment criteria.

The remainder of this article is organized as follows. The detailed of some related work in section II. And the formulation of the proposed MFFNet method is described in Section III. In Section IV, the performance of MMFNet pertaining to semantic segmentation in remote sensing image and its robustness is examined via extensive experimental demonstrations. Lastly, Section 5 presents the conclusion.

2. Related work

2.1. A. Single-Modal semantic segmentation

In recent years, deep learning technology has achieved substantial advancements in the field of remote sensing semantic segmentation. These methods can be roughly divided into three categories: (1) The pure CNN based methods [18] [19] [20]. Using a multi-context paradigm, DMS [18] presents a novel method for semantic segmentation of remote sensing images. An end-to-end trainable deep convolutional neural network for semantic segmentation is presented by DCNN [19], which combines boundary awareness with semantic significance from the outset. FactSeg [20] introduces a foreground activation (FA)-driven framework for the semantic segmentation of small objects in large-scale remote sensing images. (2) The pure Transformer based methods [21] [22] [23]. A new transformer model called Efficient Transformer [21] is presented. Its goal is to accomplish lightweight edge classification for the segmentation of remote sensing images. In order to recover resolution and produce the segmentation map, DDCST [22] presents the Swin Transformer as the foundation for collecting context information and creates a unique decoder with a densely connected feature aggregation module. CG-Swin [23] proposes a Transformer-based encoder-decoder structure for semantic segmentation of remote sensing images. (3) The CNN-Transformer based methods [24] [25] [26]. STransFuse [24] presents a new semantic segmentation method for remote sensing images that combines the advantages of Transformers and CNNs to enhance the segmentation quality across a variety of remote sensing imagery. A novel encoder-decoder structured semantic segmentation network called CMTFNet [25] is proposed to extract and fuse multiscale global contextual information with local information from high-resolution remote sensing images. A hybrid deep neural network called SwinB-CNN [26] combines Transformers and convolutional neural networks (CNNs) to separate very high resolution (VHR) remote sensing imagery semantically. A number of other remote sensing oriented efforts [55–58] have also yielded many good results. PKINet [55] introduce the Poly Kernel Inception Network to handle the background noise and sparse feature representations. And some researchers [56] proposes a nonparametric alternative based on non-learnable prototypes. PiCo [57] propose a pixel-wise contrastive algorithm for semantic segmentation in the fully supervised learning setting. Some researchers [58] address hierarchical semantic segmentation (HSS), with the aim of providing a structured, pixel-wise description of visual observation in terms of a class hierarchy

While single-modal approaches have achieved notable advancements, constraints inherent to RGB imagery hinder the extraction of intricate details from high-resolution spatial data. Consequently, incorporating multimodal strategies for remote sensing semantic segmentation bolsters feature recognition and classification precision, particularly in intricate scenarios and during adverse weather, thereby enhancing the robustness of the model. And the MMFNet model innovatively integrates multi-modal image processing techniques through a three-branch architecture. Specifically, the DSM modality provides valuable spatial and structural information, which effectively complements the RGB modality and enhances the accuracy of image segmentation. Moreover, we introduce a new Local-Global Feature Fusion module, specifically designed to efficiently combine information from both modalities by capturing both local details and global context.

2.2. B. Multi-Modal semantic segmentation

Fusing hyperspectral, multispectral, and high-resolution images, multimodal remote sensing unveils finer details like land cover [27], vegetation health [28], soil moisture [29], vital for precision agriculture [30], disaster monitoring[31], and urban planning[32]. Over the years, the application of multimodal methods [16] [15] [33] [17] has seen substantial growth. For example, CMGFNet [16] develops an optimal cross-modal fusion network that leverages the strengths of both modalities for remote sensing semantic segmentation. CIMFNet [15] devises a multimodal fusion module to investigate the similarities and differences between features from the two information modalities, facilitating effective and adequate fusion. MIMRF [33] present a novel Multiple Instance Multi-Resolution Fusion framework that can fuse multi-resolution and multi-modal sensor outputs while learning from automatically generated, imprecisely-labeled data. Above methods are based on the framework of convolutional neural networks, and some CNN-Transformer methods are described below. For examples, FTransUNet [17] presents a new fusion module called FVit for extracting semantic information and spatial relationships. A lightweight multi-modal semantic segmentation architecture based on Transformers is proposed by MFTransNet [34]. The development of these multimodal methodologies has subsequently led to the emergence of various fusion schemes. Early fusion merges spectral data (RGB, NIR) with DSM upfront, feeding this combined input to singlemodal or multimodal networks. Moreover, late fusion processes spectral and structural data through separate encoder-decoders, merging class predictions only in the final stage.

All of the above methods have achieved good outcomes, but none of them have been considered for the segmentation of small objects. The integration of multi-scale features is instrumental in improving the segmentation accuracy of small-sized objects. Moreover, the aforementioned methods fail to consider that the fused information between the RGB and DSM modalities can be leveraged to produce more accurate segmentation outputs. And MMFNet innovatively takes into account both the individual RGB and DSM modalities as well as their fused

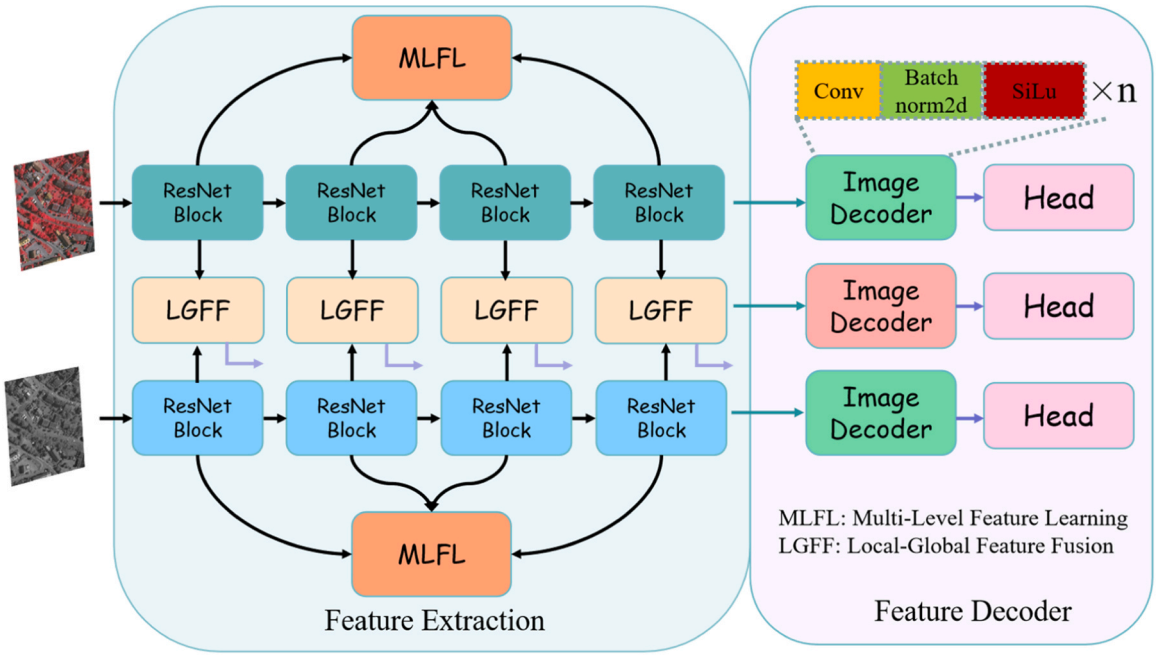

Fig. 2. Overall architecture of proposed MMFNet.

representation, leading to the development of the first three-branch network structure in this domain. This architecture effectively facilitates precise image segmentation.

2.3. C.Multi-Level feature fusion

In remote sensing images, objects are usually presented at different scales. The hierarchical architecture of Convolutional Neural Networks engenders a distribution of information where shallow, high-resolution layers harbor detail-rich and positionally accurate data, advantageous for discerning small-scale objects. Conversely, deep, low-resolution layers encapsulate high-level semantic content pivotal for recognizing larger-scale entities. Consequently, the integration of multi-scale attributes, which include both low-level, fine-grained details and high-level semantic features, is crucial for enabling models to effectively identify objects of various sizes. Recently, some work [35] [36] [37] has been successful in this field. For example, FPN [35] designs a feature pyramid network for object detection. NAS-FPN [36] adopts Neural Architecture Search and devise a new feature pyramid architecture covering all cross-scale connections. BiFPN [37] develops a weighted bi-directional feature pyramid network, which allows easy and fast multiscale feature fusion. At the same time, in recent years, some feature fusion networks [53,54] have been carefully designed and developed in the field of remote sensing image processing, and these networks have demonstrated excellent performance in a variety of applications.

Although obtaining promising results, these methods still need to be strengthened in integrating complex semantic properties of remotely sensed images for in-depth exploration. Different from them, a new multilevel feature learning (MLFL) module is designed for remote sensing semantic properties to further improve the model performance in multimodal remote sensing semantic segmentation.

2.4. D.Different level and modalities fusion module

With the advancement of multimodal approaches, the fusion of different modalities under different stages of feature extraction becomes more and more important. This evolution has prompted a surge of novel methodologies emerging from a spate of recent academic investigations. For example, FTransNet [17] devises a shallow-level feature fusion (SFF) module inspired by the SE[38] module, which allows the output of two modalities to be summed up by the attention mechanism. Moreover, in order to combine two modalities, CMGFNet [16] suggests a gated fusion module (GFM) that evaluates the usefulness of each relevant lateral feature from the RGB and DSM encoders and gathers data appropriately. Additionally, CIMFNet [15] develops a cross-layer gate fusion module (CGFM) that aims to establish dependencies to enhance features across adjacent levels. Therefore, in order to focus on both global and local information, we create a multimodal fusion module that incorporates attention mechanisms and different pooling operations. Specifically, we design the module to emphasize RGB images more than DSM data because to its richer visual feature.

3. Proposed method

Fig. 2 illustrates the process flow for extracting building information from very high resolution (VHR) remote sensing imagery and digital surface models (DSM) derived from LiDAR data using an end-to-end Multimodal Fusion Network (MMFNet) architecture. As shown in the Fig. 2., this architecture includes the many ResNet Blocks, three Image Decoder, and three Head module et al. The ResNet block is the fundamental building block of ResNet-101[9]. The image decoder consists of multiple stacked layers, including convolutional layers, BatchNorm2d layers, and SiLU activation functions. Finally, the Head module is a simple convolutional layer. The proposed network is built on a local-global feature fusion (LGFF) module to comprehensively integrate features from different modalities. Furthermore, we design a multi-level feature learning (MLFL) module to capture feature representations at various levels. Individually, the streams process the VHR image RGB bands and the singular-band DSM data, utilizing the ResNet-101[9] model as the foundational backbone in their respective encoding phases. Finally, combining single-modal and multi-modal segmentation outputs to concurrently ascertain prediction details. We will go into more detail about the main elements of this architecture in the sections that follow.

3.1. A.Two-Branch feature extraction

The MMFNet utilizes a two-branch feature extraction to learn RGB and DSM feature information. Specially, the ResNet-101 is employed in

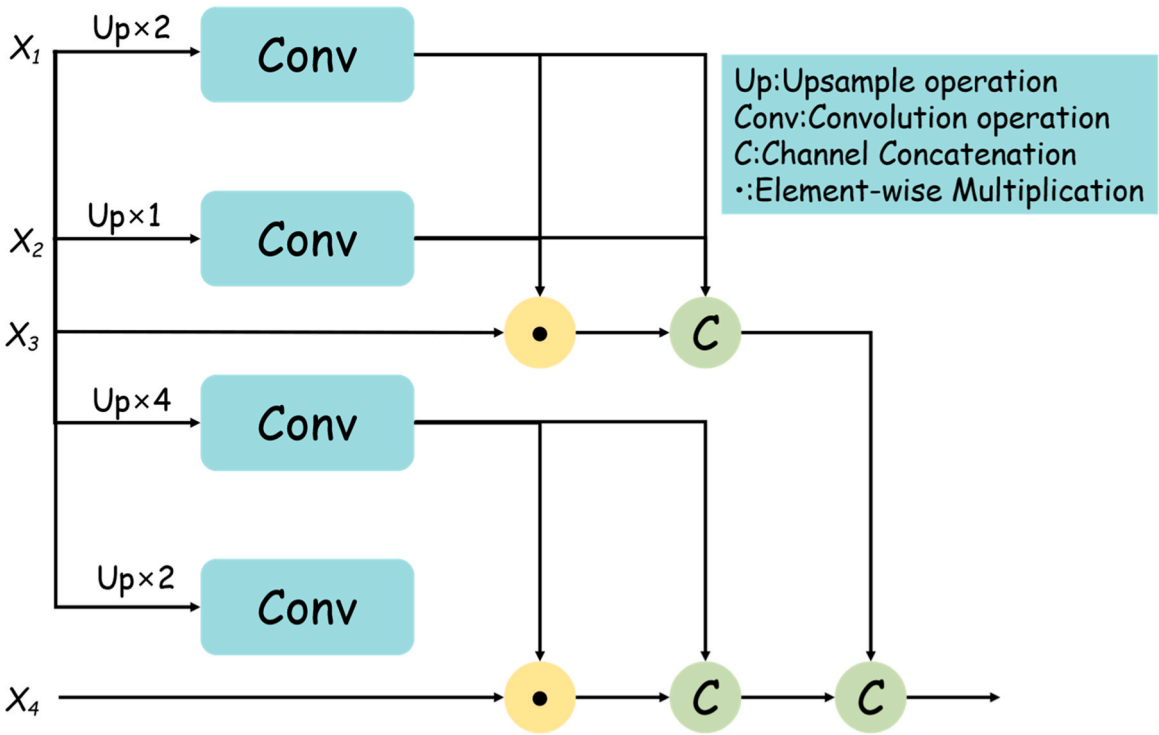

Fig. 3. Details of the introduced MLFL.

the backbone network. We first feed two images $I_1 \in R^{3\times H\times W}$ and $I_2 \in R^{1\times H\times W}$ (RGB and DSM data) to generate different multi-scale feature maps $X_{1i} \in R^{C_i\times H_i\times W_i}$ (RGB) and $X_{2i} \in R^{C_i\times H_i\times W_i}$ (DSM), where $i \in \{1,2,3,4\}$, $C_i \in \{256, 512, 1024, 2048\}$, H and W denote height and width respectively. The backbone part of the convolutional operation initializes its weights using a pre-trained model from the ImageNet dataset.

3.2. B. Multi level feature learning module

The purpose of the proposed Multi Level Feature Learning (MLFL) module is to learn multi-level feature from different stages feature maps. The detailed structure of Multi Level Feature Learning (MLFL) module is shown in Fig. 3. We first process the feature maps from the different stages $\{X_1, X_2, X_3, X_4\}$. We feed two feature maps X_1 and X_2 to get new feature maps by using upsample operation. Then we use the new feature maps to get X_{1_1} and X_{2_1} by using convolution operation. The process can be summarized as follows:

$$X_{1_1} = conv(up(X_1)) \tag{1}$$

$$X_{2_1} = conv(up(X_2)) \tag{2}$$

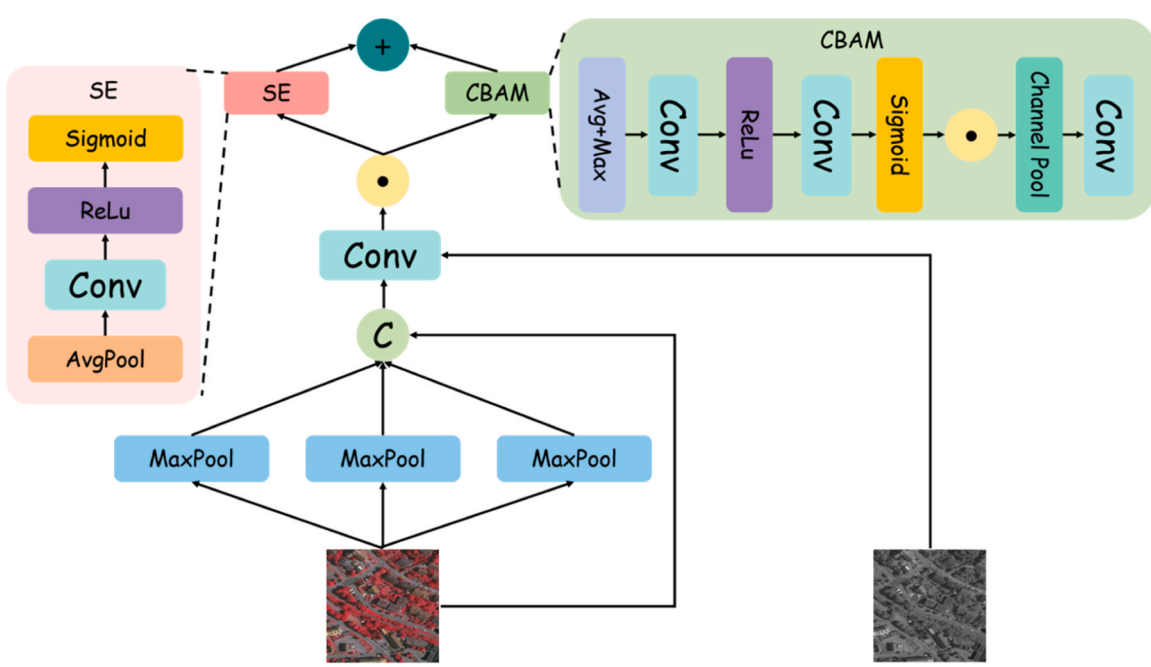

Fig. 4. A network structure with the LGFF.

Moreover, in order to generate the new feature map X_3, we plan to multiply by other feature maps X_{1_1} and X_{2_1}.Then all the feature maps are fused to get new feature map $out1$. They are formulated as follows:

$$X_3 = X_{1_1} \times X_{2_1} \times X_3 \tag{3}$$

$$out1 = conv(cat[X_{1_1}, X_{2_1}, X_3]) \tag{4}$$

Next, we will introduce the another output feature map $out2$. At the same time, in order to gengerate the new feature map X_4, we firstly feed two feature maps X_1 and X_2 to get new feature maps by using upsample operation. Then we use the new feature maps to get X_{1_2} and X_{2_2} by using convolution operation. The process can be summarized as follows:

$$X_{1_2} = conv(up(X_1)) \tag{5}$$

$$X_{2_2} = conv(up(X_2)) \tag{6}$$

Then, we plan to multiply by other feature maps X_{1_2} and X_{2_2}.Then all the feature maps are fused to get new feature map $out2$. They are formulated as follows:

$$X_4 = X_{1_2} \times X_{2_2} \times X_4 \tag{7}$$

$$out2 = conv(cat[X_{1_2}, X_{2_2}, X_4]) \tag{8}$$

Ultimately, we fuse the two produced feature maps ($out1$ and $out2$) to derive the resultant output feature map ($output$). This process is formulated as follows:

$$output = cat[out1, out2] \tag{9}$$

Where up is upsample operation, $conv$ is convolution operation and cat is channel concatenation operation. At the same time, the size of the convolution kernel is 3×3.

This study proposes an innovative multi-level feature fusion module designed to adapt to the learning needs of different modalities. The framework achieves effective feature fusion of RGB and DSM data by hierarchically integrating multilevel features extracted from multiple backbone networks. Specifically, we design a multilevel fusion module, which is capable of preserving high-resolution spatial details in the shallow layer and gradually introducing high-level semantic features in the deep layer. This progressive fusion not only effectively mitigates the differences in feature distribution among different modalities, but also enhances the model's detection sensitivity to small-scale targets, thus improving the overall system performance.

3.3. C. Local-Global feature fusion module

Compared with traditional multimodal fusion methods, most of them only consider local features without considering global features. And RGB images encapsulate a broader spectrum of visual information compared to the relatively more singular data offered by DSM images. Therefore, we design a new Local-Global Feature Fusion (LGFF) module to fuse different multimodal feature maps by using pooling operations. The structure of the LGFF is shown in Fig. 4. The output of the feature map on the RGB and DSM backbones, respectively, is represented by Frgb and Fdsm. It is important to note that every modality has the same feature dimensions.

Initially, the RGB image undergoes processing via three distinct pooling layers (f_1, f_2 and f_3), subsequent to which, the resultant features are concatenated to form a new feature representation (f_{rgb}). All pooling operations we use are max pooling layers. They are formulated as follows:

$$f_1 = pool(F_{rgb}), f_2 = pool(F_{rgb}), f_3 = pool(F_{rgb}) \tag{10}$$

In this formula, F_{rgb} represents the feature map extracted from the RGB input image. The function pool denotes a max pooling operation, which is typically used to reduce the spatial dimensions of the feature maps while retaining import information. In our model, we apply three different pooling on the same feature map. The purpose of using multiple pooling operations is to capture multi-scale features from the RGB input, which helps in improving the robustness and generalization ability of the model.

$$f_{rgb} = cat[f_1, f_2, f_3, F_{rgb}] \tag{11}$$

This formula describes the concatenation of the pooled feature maps f_1, f_2 and f_3 along with the original feature map F_{rgb}. The function cat stands for concatenation, which combines these feature maps into a new feature map f_{rgb}. By concatenating the multi-scale features obtained from different pooling operations with the original feature map, we aim to enrich the representation of the RGB input. This enriched feature map f_{rgb} contains both local and global information, which is beneficial for subsequent processing steps.

Next, we intend to multiply these two features f_{rgb} and F_{dsm}together to get a new feature map f_{fuse}. This process is formulated as follows:

$$f_{fuse} = f_{rgb} \times F_{dsm} \tag{12}$$

In this formula, f_{fuse} represents the fused feature map obtained by element-wise multiplication between the concatenated RGB feature map f_{rgb} and the DSM feature map F_{dsm}. The element-wise multiplication operation allows us to integrate the complementary information from the RGB and DSM inputs effectively. The resulting fused feature map f_{fuse} captures the combined strengths of both modalities, leading to improved performance in the remote sensing semantic segmentation task. This fusion strategy ensures that the model can leverage the rich visual cues from the RGB data and the geometric information from the DSM data simultaneously.

At the same time, in order to focus on the key areas of the feature map, we employ Squeeze and Excitation (SE) Attention [38] and Convolutional Block Attention Module (CBAM) [39] to obtain the channel-spatial attention map, which is then added together to obtain the final feature map (F_{fuse}).

$$F_{fuse} = CBAM(f_{fuse}) + SE(f_{fuse}) \tag{13}$$

Where pool denotes the global max pooling operation, $\times$ represents the multiplication operation and cat is the channel concatenation operation.

Why was the LGFF module designed this way? The SPP layer is placed at the front end, and its main function is to capture the spatial context information of the input feature map through pooling operations of different scales. The SPP layer can aggregate context information of multiple scales, which is especially important for understanding objects in complex scenes. Additionally, it allows the network to process input images of any size, as the operations following the SPP layer are fixed-sized. This design enhances the model's robustness and flexibility when handling inputs of different resolutions. The SE module aims to improve network performance by explicitly modeling the relationships between channels. It achieves this by adaptively adjusting the importance of each channel, enabling the network to focus more on features that are more important for specific tasks. This helps enhance the model's ability to selectively emphasize useful information, thereby improving overall performance. The CBAM module is a lightweight and effective attention module that can be integrated into convolutional neural networks to enhance representation capabilities. CBAM includes two independent submodules: the channel attention module and the spatial attention module. The channel attention module helps the network focus on more important feature maps, while the spatial attention module helps the network focus on more critical spatial locations. This design enables the network not only to identify which features are more important but also to pinpoint their specific spatial locations, further enhancing the model's expressive power. Overall, this design aims to achieve more effective local and global feature fusion by integrating the advantages of SPP, SE, and CBAM, thereby improving the model's sensitivity to small-scale objects and enhancing its

understanding of complex scenes.

The Local-Global Feature Fusion module we designed integrates features from different pooling layers and employs an attention mechanism to achieve efficient fusion of critical information. This approach not only effectively merges RGB and DSM data but also processes these two modalities within the same stage, thereby enhancing the quality of feature representation.

3.4. D.Loss function

With single-modal and multi-modal feature maps obtained, we have devised novel strategies for enhancing network optimization. We adopt to the binary cross-entropy (BCE) loss as our network loss function. To accelerate the convergence process, we use y_{rgb}, y_{dsm} and y_{fuse} as the prediction result of RGB, DSM image and fused-modal respectively. However, we employ y_{output} as the finally prediction result of the our network. This process is formulated as follows:

$$y_{output} = (1-\alpha)y_{fuse} + \alpha y_{rgb} \tag{14}$$

Where α denotes the weight coefficient factor.

Therefore, our loss function consists of four components: L_{rgb}, L_{dsm}, L_{fuse} and L_{output}. The loss function can be formulated as follow:

$$L_{total} = L_{rgb} + L_{dsm} + L_{fuse} + L_{output} \tag{15}$$

L_{total} represents the total sum of all losses. L_{rgb} denotes the RGB modality loss, L_{dsm} denotes the DSM modality loss, L_{fuse} represents the fused modality loss and L_{output} denotes the output loss.

For each loss function the BCE loss is defined as:

$$L_{total}(p,\widehat{p}) = L_{rgb}(p,\widehat{p}) + L_{dsm}(p,\widehat{p}) + L_{fuse}(p,\widehat{p}) + L_{otuput}(p,\widehat{p}) \tag{16}$$

And the loss function L_{total} incorporates information from all modalities, which facilitates more effective model optimization.

4. Experiments and discussion

In this section, we first introduce two datasets, evaluation metrics and experimental settings. Subsequently, MMFNet model is compared with the state-of-the-art methods. At last, we provide the detailed performance analyses and the explainable visualizations of MMFNet model.

4.1. A.Datasets

For this experiment, we utilized two publicly available ISPRS datasets: the Potsdam and Vaihingen datasets.

1. **Potsdam dataset**: There are 38 true orthophotos (TOP) in this collection, along with matching DSM patches. With a resolution of 6000 × 6000 pixels and a normalized digital surface model (DSM) at a ground sampling distance (GSD) of 5 cm, it offers four multispectral channels: InfraRed, Red, Green, and Blue (IRRGB). This dataset has one background class (Clutter) and five foreground classes: Low Vegetation (Low.), Car, Impervious Surface (Imp.), Tree (Tre.), and Building (Bui.).
2. **Vaihingen dataset:** There are sixteen true orthophotos (TOP) in this collection, along with matching DSM patches. Every orthophoto includes a normalized digital surface model (DSM) at a 9 cm GSD and a resolution of 2500 × 2000 pixels, along with three channels: Near-InfraRed, Red, and Green (NIRRG). The dataset shares one background class and five foreground classes with the Potsdam dataset.

It is important to note that our dataset division criteria are based on the guidelines provided in this article [17].

4.2. B.Evaluation metrics

The assessment of remote sensing image segmentation models typically entails meticulous evaluation of their performance and accuracy on dedicated test datasets. In the paper, we adopt to three popular metrics: overall accuracy (OA), F1-score, and mean intersection over union (mIoU). These conventional statistical metrics afford an impartial basis for comparing the performance of the proposed MMFNet with that of current state-of-the-art methodologies. Traditionally, the measure of overall accuracy (OA), or pixel-wise accuracy, is employed to assess segmentation efficacy. This index represents the proportion of correctly classified pixels relative to the total pixel count across all segments in the test datasets, computed according to the following formula:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

where *TP, TN, FP,* and *FN* denote true positive, true negative, false positive and false negative, respectively. Furthermore, F1 and mIoU are calculated for each foreground class indexed by c according to the following formula:

$$F1 = 2 \times \frac{Q_c \times R_c}{Q_c \times R_c} \tag{18}$$

$$IoU = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{19}$$

where TP_c, FP_c and FN_c are true positives, false positives, and false negatives for the c-th class, respectively. Finally, Q_c and R_c are given by:

$$Q_c = \frac{TP_c}{TP_c + FP_c} \tag{20}$$

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{21}$$

We calculate the mean values, specifically the mean F1 score (mF1) and mean Intersection over Union (mIoU), by averaging the F1 scores and IoU values for the five useful foreground classes, as defined above.

4.3. C.Implementation details

In this paper, all experiments were conducted using PyTorch on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of RAM. Following the setup in FTransUNet [17], we used Stochastic Gradient Descent (SGD) as the optimizer with a learning rate of 0.03, a momentum of 0.9, a weight decay coefficient of 0.0005, and a batch size of 16. The input images were resized to 256 × 256 pixels, and the training was performed for 50 epochs. The initial learning rate was reduced by a factor of 10 at the 25th, 35th, and 45th epochs. Additionally, to enrich the dataset, basic data augmentation techniques such as random rotations and flips were applied to the samples acquired via the sliding window approach. The time required to train one round of our model is 0.35 h and the Flops of our model is 346.18 G. All pooling operations we use are max pooling layers.

4.4. D.Experimental result

To better demonstrate the superiority of our proposed MMFNet, we have selected a number of methods for comparison, including FTransUNet [17], TransUNet [40], CMFNet [41], UNetFormer [42], MFTransNet [34], CMGFNet [16], SA-GATE [43], ESANet [44], vFuseNet [45], MAResU-Net [46], PSPNet [47], ABCNet [48], FuseNet [49], ST-UNet [14], MANet [50], DeepLabv3 +[51] and HRNet[52]. Notably, in our experiments, ABCNet, PSPNet, ST-UNet, MANet, MAResU-Net, and UNetFormer only utilized the one modal information, highlighting the importance of incorporating DSM data. Compared to these approaches, our methodology achieved the most superior performance on

Table 1
Comparison of segmentation results on the vaihingen data set.

Method	OA(%)						mF1(%)	mIoU(%)
	Bui.	Tre.	Low.	Car.	Imp.	Total		
FTransUNet[17]	96.69	92.03	79.54	84.00	92.54	91.45	89.46	81.35
TransUNet[40]	96.48	**92.77**	76.14	69.56	91.66	90.96	87.34	78.26
CMFNet[41]	94.71	91.22	78.62	85.06	92.67	90.53	89.08	80.72
UNETFormer[42]	96.23	91.85	79.95	86.99	91.85	91.17	89.85	81.97
MFTransNet[34]	96.41	91.48	80.09	86.52	92.11	91.22	89.62	81.61
CMGFNet[16]	97.75	91.60	80.03	87.28	92.35	91.72	90.00	82.26
SA-GATE[43]	94.84	92.56	81.29	87.79	91.69	91.10	89.81	81.27
ESANet[44]	95.69	90.50	77.16	85.46	91.39	90.61	88.18	79.42
vFuseNet[45]	95.92	91.36	77.64	76.06	91.85	90.49	87.89	78.92
MAResU-Net[46]	94.84	89.99	79.09	85.89	92.19	90.17	88.54	79.89
PSPNet[47]	94.52	90.17	78.84	79.22	92.03	89.94	86.55	76.96
ABCNet[48]	94.10	90.81	78.53	64.12	89.70	89.25	85.34	75.20
FuseNet[49]	96.28	90.28	78.98	81.37	91.66	90.51	87.71	78.71
ST-UNet[14]	93.44	92.37	76.44	82.83	91.71	89.79	87.92	78.86
MANet[50]	96.55	91.85	79.01	**90.08**	92.26	91.25	90.01	82.28
DeepLabv3 +[51]	94.64	92.24	76.74	83.85	90.93	89.99	88.32	79.54
HRNet[52]	95.72	91.16	77.02	74.69	90.81	89.97	87.03	77.62
MMFNet(our)	**98.10**	91.82	**81.81**	89.49	**93.13**	**92.41**	**91.14**	**84.12**

Table 2
Comparison of segmentation results on the potsdam data set.

Method	OA(%)						mF1(%)	mIoU(%)
	Bui.	Tre.	Low.	Car.	Imp.	Total		
FTransUNet[17]	97.79	88.80	86.64	96.16	92.63	90.92	91.99	85.48
TransUNet[40]	96.63	82.65	89.98	93.17	91.93	90.01	90.97	83.74
CMFNet[41]	97.54	85.16	86.24	95.53	92.13	89.81	90.98	83.81
UNETFormer[42]	97.69	86.47	87.93	95.91	92.27	90.65	91.71	85.05
MFTransNet[34]	97.37	85.71	86.92	96.05	92.45	89.96	91.11	84.04
CMGFNet[16]	97.41	86.80	86.68	95.68	92.60	90.21	91.40	84.53
SA-GATE[43]	96.54	81.18	85.35	96.63	90.77	87.91	90.26	82.53
ESANet[44]	97.10	85.31	87.81	94.08	92.76	89.74	91.22	84.15
vFuseNet[45]	97.23	84.29	89.03	95.49	91.62	90.22	91.26	84.26
MAResU-Net[46]	96.82	83.97	87.70	95.88	92.19	89.82	90.86	83.61
PSPNet[47]	97.03	83.13	85.67	88.81	90.91	88.67	88.92	80.36
ABCNet[48]	96.23	78.92	86.40	92.92	88.90	87.52	88.14	79.26
FuseNet[49]	97.48	85.14	87.31	96.10	92.64	90.58	91.60	84.86
ST-UNet[14]	94.79	80.37	84.16	91.42	88.15	85.76	87.21	77.68
MANet[50]	97.82	87.22	87.17	96.35	92.19	90.81	91.97	85.46
MMFNet(our)	98.42	85.19	89.74	95.87	**93.50**	**91.46**	**92.34**	**86.13**

remote sensing datasets.

1) Performance Comparison on the Vaihingen dataset: Table 1 displays the experimental outcomes of the various techniques on the ISPRS Vaihingen dataset. Notably, our study has achieved the apex of state-of-the-art (SOTA) performance on this dataset, confirming that our proposed MMFNet is effective in fusing different modalities and benefits from multi-level feature representation. MMFNet fared better in three classes—*Building, Low Vegetation, and Impervious Surface*—than popular state-of-the-art models. In particular, MMFNet, which is based on the multi-level feature learning (MLFL) module, demonstrated its improved segmentation accuracy for tiny objects by displaying a significant improvement of 6.66 % for the Car class when compared to the current approach ST-UNet. The classification accuracy for the Building class improved by 1.41 % compared to FTransUNet, and the classification accuracy for the Tree class increased by 1.54 % compared to FuseNet. Furthermore, compared to MFTransNet, the classification accuracy for Low Vegetation increased by 1.72 %, while compared to TransUNet, the classification accuracy for Impervious Surface increased by 1.47 %. These improvements can be attributed to MMFNet's ability to more effectively extract and fuse multi-level, multi-modal features by learning multi-scale feature maps, whereas other methods primarily focus on local features and do not fully consider global information. Multimodal image data, as opposed to single-modal data, is more advantageous for remote sensing semantic segmentation since MMFNet is made to efficiently extract and utilize global information from both RGB and DSM images. When it comes to overall effectiveness, the proposed MMFNet achieved an Overall Accuracy (OA) of 92.41 %, a mean F1 score (mF1) of 91.14 %, and a mean Intersection over Union (mIoU) of 84.12 %, representing increases of 2.62 %, 3.22 %, and 5.26 %, respectively, compared to the corresponding performance of ST-UNet. These results confirm that the proposed MMFNet achieves better generalization performance.
2) Performance Comparison on the Potsdam dataset: Excellent performance is also demonstrated by experiments on the Potsdam dataset. *Building, Tree, Low Vegetation, Car,* and *Impervious Surface* had classification accuracy rates of 98.42 %, 85.19 %, 89.74 %, 95.87 %, and 93.50 %, respectively, as shown in Table 2. These percentage increases were compared to ST-UNet at 3.63 %, 4.82 %, 5.58 %, 4.45 %, and 5.35 %. The corresponding Overall Accuracy (OA), mean F1 score (mF1), and mean Intersection over Union (mIoU) were 91.46 %, 92.34 %, and 86.13 %, respectively, which are improvements of 5.7 %, 5.13 %, and 8.45 % compared to ST-UNet. Compared with existing state-of-the-art models, our proposed MMFNet achieves the most superior state-of-the-art (SOTA) performance, particularly in the Impervious Surface class. Notably, our model performs exceptionally well on the Tree and Car classes,

Table 3
Results of ablation study on the vaihingen data set.

MLFL	LGFF	OA(%)	mF1(%)	mIoU(%)
√		91.76	89.87	82.03
	√	92.32	90.95	83.79
√	√	92.41	91.14	84.12

Table 4
Key parameter α of ablation study on the vaihingen data set.

α	OA(%)	mF1(%)	mIoU(%)
0.0	91.50	89.95	82.19
0.3	**92.41**	**91.14**	**84.12**
0.6	91.71	90.19	82.56
1.0	90.87	88.01	79.17

demonstrating the effectiveness of our proposed multi-level feature learning (MLFL) module. Additionally, MMFNet outperforms single-modal methods, highlighting the superiority of our proposed local-global feature fusion (LGFF) module. In conclusion, these results show that our proposed algorithm, MMFNet, possesses strong robustness and generalization capabilities, providing a new approach for multimodal remote sensing semantic segmentation.

4.5. E.Ablation study

Ablation studies were carried out by eliminating particular modules from MMFNet while preserving the general single-modal and multi-modal fusion architecture in order to confirm the efficacy of each suggested module. Two ablation experiments were created using our suggested network, as Table 3 demonstrates. In the first experiment, we investigate the effectiveness of the MLFL module. Moreover, we test the effectiveness of the LGFF module in the second experiment. We can see the effectiveness of the LGFF module from the first experiment, which amounts to an increase of 0.65 %, 1.27 % and 2.09 % in OA(%), mF1(%) and mIoU(%). In particular, we also can see the effectiveness of the MLFL module from the second experiment, which amounts to an increase of 0.09 %, 0.19 % and 0.33 % in OA(%), mF1(%) and mIoU(%). These experimental results reveal the efficacy of each proposed module, with LGFF exhibiting the most substantial performance boost. The experimental results above show that the MLFL module and LGFF module we designed can effectively improve the overall performance of the model through multi-scale feature fusion. The multi-scale fusion strategy can more comprehensively integrate feature information from different levels to achieve complementary advantages. This is because in the deep feature extraction process of convolutional networks, some detailed information occupying a small pixel area (such as small targets or edge areas) is easily ignored or even lost, thereby affecting the final segmentation effect. By introducing a multi-scale fusion mechanism, these feature information from low-pixel-density regions can be effectively retained, enhancing the model's ability to perceive details and significantly improving segmentation accuracy and model robustness. In summary, the design of the MLFL and LGFF modules not only enhances the model's ability to express multi-scale features but also effectively mitigates the issue of detail information loss in deep networks, providing a robust foundation for improving overall segmentation performance.

To further verify the effectiveness of the proposed model, we conducted additional ablation experiments. By varying the key parameter α across different settings, namely 0, 0.3, 0.6, and 1.0, we evaluate its impact on model performance. As shown in Table 4 , we repopulated the image segmentation experiments under different parameter settings. The experimental results show that the model performs best when the parameter α is set to 0.3. This result suggests that the segmentation performance of the model can be effectively improved by fusing appropriate multimodal information. The experimental results indicate that when the parameter is set to 0.3, the model achieves optimal overall performance. Under this parameter configuration, we expect that fused information will play a dominant role in the model's decision-making process, while RGB information will serve as an auxiliary role, used to supplement and optimise the expression of details during the fusion process. The superior performance observed when the parameter is set to 0.3 may be attributed to the model achieving an optimal balance between RGB information and fused information under this setting. This balance enables each modality to collaborate in performing its specific functions, ultimately enhancing the model's overall segmentation accuracy and robustness.

Table 5
The different fusion module of ablation study on the vaihingen data set.

Module	OA(%)	mF1(%)	mIoU(%)
MCFM	91.81	90.37	82.87
STMSF	91.39	88.78	80.34
MF-Mamba	91.65	90.26	82.67
Our	**92.41**	**91.14**	**84.12**

Table 6
The different branch of ablation study on the vaihingen data set.

Module	OA(%)	mF1(%)	mIoU(%)
RGB	90.99	88.03	78.20
Depth	83.08	74.70	62.29
RGB & Depth	91.41	88.49	79.93
Our	**92.41**	**91.14**	**84.12**

Table 7
Comparison of the complexities of different methods.

Method	Params (Mb)	OA(%)	mF1(%)	mIoU(%)
FTransUNet[17]	179.4	91.45	89.46	81.35
TransUNet[40]	32.27	90.96	87.34	78.26
CMFNet[41]	9.8	90.53	89.08	80.72
UNETFormer[42]	6.04	91.17	89.85	81.97
MFTransNet[34]	8.44	91.22	89.62	81.61
CMGFNet[16]	19.51	91.72	90.00	82.26
SA-GATE[43]	41.28	91.10	89.81	81.27
ESANet[44]	7.73	90.61	88.18	79.42
vFuseNet[45]	60.36	90.49	87.89	78.92
MAResU-Net[46]	8.79	90.17	88.54	79.89
PSPNet[47]	49.03	89.94	86.55	76.96
ABCNet[48]	3.9	89.25	85.34	75.20
FuseNet[49]	58.37	90.51	87.71	78.71
ST-UNet[14]	160.97	89.79	87.92	78.86
MANet[50]	**3.59**	91.25	90.01	82.28
MMFNet(our)	290.3	**92.41**	**91.14**	**84.12**

In order to further validate the effectiveness of the feature fusion module designed in this paper, we selected several recently proposed advanced feature fusion methods for comparison and conducted a systematic comparison and analysis under the same experimental conditions. As shown in Table 5, we selected MCFM[59], STMSF[60], MF-Mamba[61], and other feature fusion methods for comparison experiments. The results show that the feature fusion strategy proposed in this paper outperforms other fusion schemes in terms of performance. The experiment further verified that the MLFL module we proposed has good efficiency. The multi-level fusion strategy adopted by this module can effectively integrate feature information from different levels, realise cross-level feature complementarity and enhancement, thereby improving the model's perception ability and overall performance for multi-scale targets.

To further investigate the advantages of the three-branch architecture, we augmented our experimentation with additional tests. Specifically, we employed the tri-branch configuration for output generation, as detailed in Table 6. This setup was then benchmarked against single-

Table 8
Comparison of the complexities of different backbone.

Backbone	Inference Time(s)	GFLOPs (G)	Params (Mb)	OA (%)	mF1 (%)	mIoU (%)
ResNet−34	11.3	42.3	97.99	91.07	89.29	81.13
ResNet−50	11.9	54.2	116.8	91.49	90.44	82.92
ResNet−101	12.8	135.4	290.3	**92.41**	**91.14**	**84.12**
ResNet−152	15.6	294.94	438.29	91.41	88.50	79.95

branch and dual-branch configurations to discern comparative performance. The empirical evidence gathered from these experiments unequivocally demonstrates that our proposed three-branch structure yields superior outcomes. This enhanced performance underscores the efficacy of integrating multiple branches within the architecture, which facilitates more nuanced and robust feature extraction, ultimately contributing to improved overall model accuracy and reliability.

4.6. F.Model complexities

In Table 7, we analyze the parameters of the MMFNet model and compare it with various remote sensing semantic segmentation methods. Although our model has a higher intrinsic complexity, this is due to the integration of a deeper backbone feature extraction network and a three-branch architecture. Compared to lightweight networks such as SA-GATE, vFuseNet, and ST-UNet, the proposed MMFNet delivers some improvements in segmentation performance, despite a moderate increase in model complexity.

To further investigate the relationship between model complexity and performance, we designed a series of comparative experiments. Specifically, we selected different variants of the ResNet architecture, including ResNet-34, ResNet-50, ResNet-101, and ResNet-152. Our experimental results demonstrate that increased model complexity does not necessarily lead to improved performance. For instance, both the ResNet-34 to ResNet-50 and the ResNet-50 to ResNet-101 transitions show performance improvements alongside increasing model complexity.

Moreover, using the ResNet-50 backbone, our method achieves state-of-the-art (SOTA) performance as shown in Table 2 when compared with various existing algorithms. However, when employing the larger ResNet-152 architecture, the model's accuracy begins to degrade. This observation indicates that increasing model complexity does not always result in higher accuracy, and the correlation between the two is not strictly positive.

As shown in Table 8, we further supplement the GLOPs and inference time for each model version. The inference time refers to the total time required to perform inference on the entire test set. Based on the performance of the aforementioned models, they are well-suited for local deployment in practical applications.

4.7. G.Visualization and analyses

In this section, we have deliberately chosen a selection of algorithms, including CMFNet, ST-UNet, MANet, FTransUNet, and MFFNet, for visual representation and comparative analysis alongside our proposed algorithm. This endeavor facilitates a comprehensive evaluation, highlighting the distinctive features and performance characteristics of each model in relation to our innovative approach. We selected the test dataset from the Vaihingen dataset for visualization. As shown in Fig. 5., our proposed algorithm demonstrably excels in segmentation performance, exhibiting a heightened level of robustness. This innovative multimodal remote sensing semantic segmentation methodology introduces fresh perspectives and concepts to the field.

5. Conclusion

In this work, the end-to-end multimodal fusion network called MMFNet has been proposed for remote sensing semantic segmentation by utilizing local-global feature fusion and multi level feature learning. This work pioneers the integration of a novel three-branch multimodal framework into remote sensing semantic segmentation, uniquely

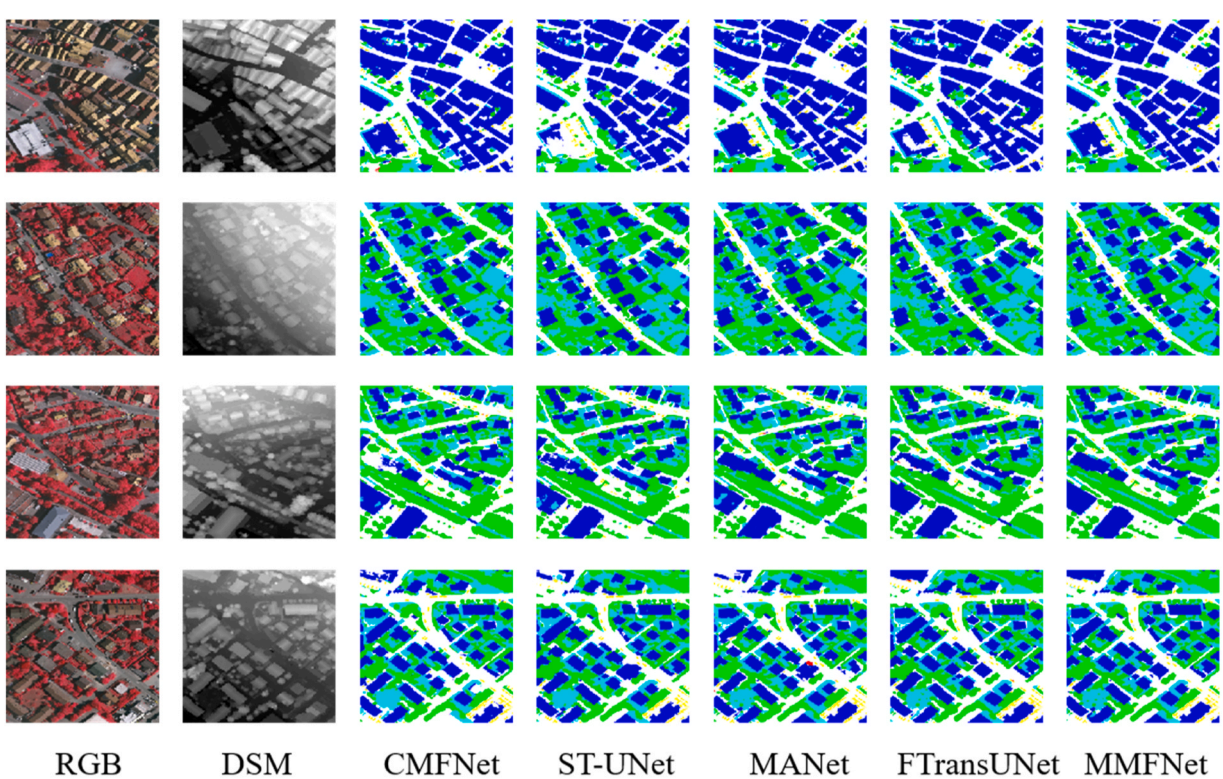

Fig. 5. Qualitative comparison on Vaihingen datasets.

merging both single-modal and multi-modal output information. It marks the first instance where such an intricate design has been applied to this domain, thereby signifying a groundbreaking contribution. Comprehensive evaluations conducted on two prominent datasets - ISPRS Vaihingen and Potsdam - have decisively validated that our proposed MMFNet surpasses current state-of-the-art segmentation techniques in terms of performance, affirming its advanced capabilities. In conclusion, acknowledging the substantial complexity inherent in our model, we aspire, in future endeavors, to undertake research directed at developing a more streamlined, lightweight variant. This will not only enhance computational efficiency but also broaden the practical applicability of our methodology. At the same time, we aim to extend the application of our model to other domains, such as object detection, scene classification, and more.

CRediT authorship contribution statement

Pingzhu Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Innovation and Entrepreneurship Training Program for College Students project grant number 202310410022

Data availability

Data will be made available on request.

REFERENCES

[1] S. Ying, S. Zhou, X. Jiang, Y. Gao, F. Pan, Z. Fang, High-resolution enhanced cross-subspace fusion network for light field image superresolution, Displays 84 (2024).
[2] Q. Zhu, Y. Zhang, L. Wang, Y. Zhong, Q. Guan, X. Lu, L. Zhang, D. Li, A global context-aware and batch-independent network for road extraction from VHR satellite imagery, ISPRS J. Photogramm. Remote Sens. 175 (2021) 353–365.
[3] L. Wang, X. Wang, W. Liu, H. Ding, B. Xia, Z. Zhang, J. Zhang, S. Xu, A unified architecture for super-resolution and segmentation of remote sensing images based on similarity feature fusion, Displays 84 (2024).
[4] K. Zhou, D. Ming, X. Lv, J. Fang, M. Wang, CNN-Based land cover classification combining stratified segmentation and fusion of point cloud and very high-spatial resolution remote sensing image data, Remote Sens. (2019).
[5] H. Zhu, L. Li, J. Wu, W. Dong, G. Shi, MetaIQA: deep Meta-Learning for No-Reference image quality assessment, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR (2020) 14143–14152.
[6] C. Zhang, R. Wang, J.-W. Chen, W. Li, C. Huo, Y. Niu, A Multi-Branch U-Net for water area segmentation with Multi-Modality remote sensing images, IGARSS 2023 2023 IEEE Int. Geosci. Remote Sens. Symp. (2023) 5443–5446.
[7] F. Yuan, L. Zhang, X. Xia, Q. Huang, X. Li, A gated recurrent network with dual classification assistance for smoke semantic segmentation, IEEE Trans. Image Process. 30 (2021) 4409–4422.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. (NeurIPS) (2017) 5998–6008.
[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for imag recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2016) 770–778.
[10] L. Yuan, Q. Hou, Z. Jiang, J. Feng, S. Yan, VOLO: vision outlooker for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (5) (2023) 6575–6586.
[11] L. Zhu, B. Liao, Z. Qian, X. Wang, W. Liu and X. Wang,"Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model," arXiv: 2401.09417, 2024.
[12] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional Encoder-Decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[13] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, 234–24, Proc. Conf. Med. Image Comput. Comput. Assist. Interv. (2015), 234–24.
[14] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, Y. Xue, Swin transformer embedding UNet for remote sensing image semantic segmentation, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.
[15] W. Zhou, J. Jin, J. Lei, L. Yu, CIMFNet: Cross-Layer interaction and multiscale fusion network for semantic segmentation of High-Resolution remote sensing images, IEEE J. Sel. Top. Signal Process. 16 (4) (2022) 666–676.
[16] H. Hosseinpour, F. Samadzadegan, F. Javan, CMGFNet: a deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images, ISPRS J. Photogramm. Remote Sens. 184 (2021) 96–115.
[17] X. Ma, X. Zhang, M.-O. Pun, M. Liu, A multilevel multimodal fusion transformer for remote sensing semantic segmentation, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–15.
[18] K. Nogueira, M. Dalla Mura, J. Chanussot, W.R. Schwartz, J.A. dos Santos, Dynamic multicontext segmentation of remote sensing images based on convolutional networks, IEEE Trans. Geosci. Remote Sens. 57 (2019) 7503–7520.
[19] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu and U. Stilla, "Classification With an Edge: Improving Semantic Image Segmentation with Boundary Detection," arXiv:1612.01337, 2016.
[20] A. Ma, J. Wang, Y. Zhong, Z. Zheng, FactSeg: foreground Activation-Driven small object semantic segmentation in Large-Scale remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16.
[21] Z. Xu, W. Zhang, T. Zhang, Z. Yang, J. Li, Efficient transformer for remote sensing image segmentation, Remote Sens. (2021).
[22] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, S. Fang, A novel transformer based semantic segmentation scheme for Fine-Resolution remote sensing images, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[23] X. Meng, Y. Yang, L. Wang, T. Wang, R. Li, C. Zhang, Class-Guided swin transformer for semantic segmentation of remote sensing imagery, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[24] L. Gao, et al., STransFuse: fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 10990–11003.
[25] H. Wu, P. Huang, M. Zhang, W. Tang, X. Yu, CMTFNet: CNN and multiscale transformer fusion network for Remote-Sensing image semantic segmentation, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12.
[26] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, C. Wang, Transformer and CNN hybrid deep neural network for semantic segmentation of Very-High-Resolution remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–20.
[27] D. Hong, J. Hu, J. Yao, J. Chanussot, X. Zhu, Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model, ISPRS J. Photogramm. Remote Sens. 178 (2021) 68–80.
[28] R. Giovos, D. Tassopoulos, D. Kalivas, N. Lougkos, A. Priovolou, Remote sensing vegetation indices in viticulture: a critical review, Agriculture (2021).
[29] J. Dong, W.T. Crow, K.J. Tobin, M.H. Cosh, D.D. Bosch, P.J. Starks, M. Seyfried, C. Holifield Collins, Comparison of microwave remote sensing and land surface modeling for surface soil moisture climatology estimation, Remote Sens. Environ. 242 (2020).
[30] R.P. Sishodia, R.L. Ray, S.K. Singh, Applications of remote sensing in precision agriculture: a review, Remote Sens. (2020).
[31] M. Kucharczyk, C.H. Hugenholtz, Remote sensing of natural hazard-related disasters with small drones: global trends, biases, and research opportunities, Remote Sens. Environ. 264 (2021).
[32] T. Wellmann, A. Lausch, E. Andersson, S. Knapp, C. Cortinovis, J. Jache, S. Scheuer, P. Kremer, A. Mascarenhas, R. Kraemer, At Haase, F. Schug, D. Haase, Remote sensing in urban planning: contributions towards ecologically sound policies, Landsc. Urban Plan. 204 (2020).
[33] X. Du, A. Zare, Multiresolution multimodal sensor fusion for remote sensing data with label uncertainty, IEEE Trans. Geosci. Remote Sens. 58 (4) (2020) 2755–2769.
[34] S. He, H. Yang, X. Zhang, X. Li, MFTransNet: a multi-modal fusion with cnn-transformer network for semantic segmentation of HSR remote sensing images, Mathematics 11 (3) (2023) 722.
[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2017) 936–944.
[36] G. Ghiasi, T.-Y. Lin, Q.V. Le, NAS-FPN: learning scalable feature pyramid architecture for object detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2019) 7029–7038.
[37] M. Tan, R. Pang, Q.V. Le, EfficientDet: scalable and efficient object detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR (2020) 10778–10787.
[38] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation networks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR (2018) 7132–7141.
[39] S. Woo, J. Park, J. Lee, I. Kweon, CBAM: convolutional block attention module, Proc. Eur. Conf. Comput. Vis. (ECVV (2018) 3–19.
[40] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L.Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.
[41] X. Ma, X. Zhang, M.-O. Pun, A crossmodal multiscale fusion network for semantic segmentation of remote sensing data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15 (2022) 3463–3474.
[42] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, P.M. Atkinson, UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, ISPRS J. Photogramm. Remote Sens. 190 (2022) 196–214.

[43] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semanticsegmentation, Eur. Conf. Comput. Vis. (ECCV) (2020) 561–577.
[44] D. Seichter, M. Kohler, B. Lewandowski, T. Wengefeld, H.-M. Gross, Efficient rgb-d semantic segmentation for indoor scene analysis, 2021 IEEE Int. Conf. Robot. Autom. (ICRA (2021) 13525–13531.
[45] N. Audebert, B. Le Saux, S. Lef̀evre, Beyond RGB:Very high resolution urban remote sensing with multimodal deep networks, ISPRS J. Photogramm. Remote Sens. 140 (2018) 20–32.
[46] R. Li, S. Zheng, C. Duan, J. Su, C. Zhang, Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[47] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017) 2881–2890.
[48] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, P.M. Atkinson, ABCNet: attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery, ISPRS J. Photogramm. Remote Sens. 181 (2021) 84–98.
[49] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, FuseNet:I ncorporating depth into semantic segmentation via fusionbased cnn architecture, Asian Conf. Comput. Vis. (2016) 213–228.
[50] R. Li, et al., "Multiattention network for semantic segmentation of Fine-Resolution remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.
[51] C. Liang Chieh, et al., "Rethinking atrous convolution for semantic image segmentation, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR (2017).
[52] K. Sun, et al., "Deep High-Resolution representation learning for human pose estimation, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2019).
[53] L. Jiaxin, et al., "Deep learning in multimodal remote sensing data fusion: a comprehensive review, Int. J. Appl. Earth Obs. Geoinf. 112 (2022).
[54] Z. Zhao, et al., "Building extraction from High-Resolution multispectral and Sar images using a Boundary-Link multimodal fusion network, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 18 (2025) 3864–3878, https://doi.org/10.1109/JSTARS.2025.3525709.
[55] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, Y. Yao, "Poly kernel inception network for remote sensing detection, 2024 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Seattle WA USA (2024) 27706–27716, https://doi.org/10.1109/CVPR52733.2024.02617.
[56] T. Zhou, W. Wang, E. Konukoglu, L. Van Goo, "Rethinking semantic segmentation: a prototype view, 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) N. Orleans LA USA (2022) 2572–2583, https://doi.org/10.1109/CVPR52688.2022.00261.
[57] T. Zhou, W. Wang, "Cross-Image pixel contrasting for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 46 (8) (Aug. 2024) 5398–5412, https://doi.org/10.1109/TPAMI.2024.3367952.
[58] L. Li, W. Wang, T. Zhou, R. Quan, Y. Yang, "Semantic Hierarchy-Aware segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 46 (4) (April 2024) 2123–2138, https://doi.org/10.1109/TPAMI.2023.3332435.
[59] Z. Li, K. Yan, C. Wang, D. Zhou, "A multi-branch and multi-scale feature fusion underwater image enhancement algorithm for diverse water environments, Appl. Soft Comput. 180 (2025).
[60] Y. Duan, C. Song, Y. Zhang, P. Cheng, S. Mei, STMSF: swin transformer with Multi-Scale fusion for remote sensing scene classification, Remote Sens 17 (2025) 668.
[61] G. Li, Q. Huang, W. Wang, L. Liu, Selective and multi-scale fusion mamba for medical image segmentation, Expert Syst. Appl. 261 (2025).

Pingzhu Liu The author graduated from Jiangxi Agricultural University with a bachelor's degree and is currently pursuing a master's degree at Tsinghua University Shenzhen Graduate School.