MMFNet: A Mamba-Based Multimodal Fusion Network for Semantic Segmentation of Remote Sensing

Jingting Qiu , Wei Chang * , Wei Ren , Shanshan Hou , Ronghao Yang *

Posted Date: 4 August 2025

doi: 10.20944/preprints202508.0078.v1

Keywords: multimodal semantic segmentation; remote sensing; feature fusion

 doi:10.20944/preprints202508.0078.v1

Article

MMFNet: A Mamba-Based Multimodal Fusion Network for Semantic Segmentation of Remote Sensing

Jingting Qiu [1], Wei Chang [2,3,*], Wei Ren [2,3], Shanshan Hou [2,3] and Ronghao Yang [1,*]

[1] College of Earth and Planetary Sciences, Chengdu University of Technology, Chengdu 610059, China
[2] Qi-Liang Spatiotemporal Information Innovation Studio, the 9th Geological Brigade of Sichuan Province, Deyang 618029, China
[3] Geographic Information and Geological Environment Key Laboratory of DeyangCity, Deyang 618029, China
* Correspondence: chw0077@163.com (W.C.); yangronghao@cdut.edu.cn (R.Y.)

Abstract

Accurate semantic segmentation of high-resolution remote sensing imagery is challenged by substantial intra-class variability, inter-class similarity, and the limitations of single-modality data. This paper proposes MMFNet, a novel multimodal fusion network that leverages the Mamba architecture to efficiently capture long-range dependencies for semantic segmentation tasks. MMFNet adopts a dual-encoder design, combining ResNet-18 for local detail extraction and VMamba for global contextual modeling, striking a balance between segmentation accuracy and computational efficiency. A Multimodal Feature Fusion Block (MFFB) is introduced to progressively integrate complementary information from optical imagery and digital surface models (DSMs) via multi-kernel convolution and window-based cross-attention. Furthermore, a frequency-aware upsampling module (FreqFusion) is incorporated in the decoder to enhance boundary delineation and recover fine spatial details. Extensive experiments on the ISPRS Vaihingen and Potsdam benchmarks demonstrate that MMFNet achieves mean IoU scores of 83.50% and 86.06%, outperforming eight state-of-the-art methods while maintaining relatively low computational complexity. These results highlight MMFNet's potential for efficient and accurate multimodal semantic segmentation in remote sensing applications.

Keywords: multimodal semantic segmentation; remote sensing; feature fusion

1. Introduction

Remote sensing images (RSIs) serve as a vital data source for Earth observation. Semantic segmentation of RSIs, also referred to as land use and land cover (LULC) classification [1], aims to assign a semantic label to each pixel, enabling pixel-level categorization. LULC classification plays a crucial role in various applications, including urban planning [2], agricultural management [3] environmental monitoring [4], and the development of geographic information system (GIS) [5].

With advances in sensor technology, the spatial resolution of RSI has significantly improved. High-resolution RSIs (HRRSIs) contain rich texture, geometric, and spectral information [6], but also exhibit considerable diversity and complexity. These images are characterized by large intra-class variance, small inter-class variance, and low class separability [7], posing significant challenges to semantic segmentation tasks—for instance, the difficulty in distinguishing between trees and low vegetation in the first row of Figure 1. Moreover, shadows in HRRSIs further complicate feature extraction, thereby reducing segmentation accuracy. An example is shown in the second row of Figure 1, where shadowing impedes the identification of a path between tree canopies [8].

 doi:10.20944/preprints202508.0078.v1

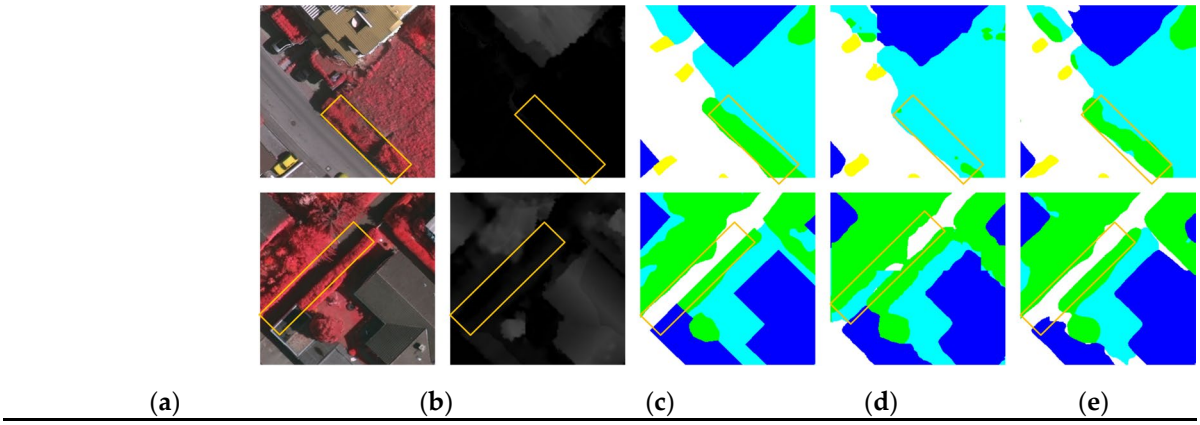


Figure 1. Predictions for objects with similar color and texture. (**a**) RGB image; (**b**) Digital Surface Model (DSM); (**c**) Ground truth (GT); (**d**) Prediction using unimodal data; (**e**) Prediction using multimodal data.

Compared with unimodal approaches, multimodal semantic segmentation (MSS) integrates diverse remote sensing sources—such as digital surface models (DSMs), multispectral, panchromatic, and radar data—to exploit complementary information across modalities, thereby improving segmentation accuracy [9]. For instance, DSMFNet [10] introduces a DSMF module that extracts informative features from DSM data to enhance segmentation performance in shadowed regions or areas with similar colour and texture. CMGFNet [11] employs a Gated Fusion Module (GFM) to adaptively combine features from different modalities while suppressing redundancy. Similarly, CMFNet [12] uses a cross-attention mechanism to fuse multimodal features at multiple scales, achieving cross-modal integration with multiscale awareness.

Most existing multimodal segmentation networks, however, rely on convolutional neural networks (CNNs) [13] or vision transformer (ViT) [14] as backbone architectures. CNNs excel at extracting local features due to their local connectivity and weight sharing, but their limited receptive field constrains their ability to model global context [15]. ViT [14], on the other hand, offer stronger long-range dependency modelling, but their quadratic computational complexity hinders their applicability to high-resolution imagery [16]. Thus, multimodal segmentation models based on CNNs or ViT face a fundamental trade-off between capturing global context and maintaining computational efficiency.

Mamba [17], a variant of the State Space Model (SSM), was originally introduced for natural language processing. It offers the capability to model long-range dependencies while maintaining linear computational complexity. Its successful adaptation to RSI segmentation, as demonstrated by Vim [18] and VMamba [19], highlights its potential as an alternative to Transformer-based architectures. Sigma [20] represents the first attempt to apply SSMs to multimodal segmentation, proposing an end-to-end network entirely built upon the Mamba architecture. Specifically, Sigma adopts VMamba as the feature extraction backbone and designs two Mamba-based fusion modules, Cross Mamba Block (CroMB) and Concat Mamba Block (ConMB), to facilitate cross-modal interaction. Despite Mamba's strength in capturing long-range dependencies, it performs suboptimally in segmenting small-scale objects. To address this limitation, MFMamba [21] combines CNN and VMamba backbones to separately extract multimodal features. It incorporates a Feature Fusion Block (FFB) to enhance both global and local modality-specific features, which are then aggregated via element-wise addition. While this approach effectively captures multi-scale global and local features, it remains susceptible to information loss and feature misalignment, particularly in complex remote sensing scenes with large inter-modal differences.

In summary, several challenges persist in MSS of RSI:

(1) Balancing accuracy and efficiency. MSS is a dense, pixel-level classification task. Current approaches often adopt dual-encoder frameworks based on CNNs and Transformers to capture local and global semantic features, respectively. However, Transformer-based methods incur substantial

 doi:10.20944/preprints202508.0078.v1

computational overhead, making it difficult to achieve both high accuracy and computational efficiency.

(2) Effective multimodal feature fusion. Most existing methods perform feature fusion through concatenation or element-wise addition. However, when significant differences exist between modalities, such strategies may lead to information loss and feature misalignment. Moreover, redundant information may be introduced, amplifying the impact of image noise and ultimately degrading segmentation accuracy [22].

To address these challenges, we propose MMFNet, a multimodal dual-encoder fusion network based on the Mamba architecture. In this framework, The primary encoder adopts ResNet-18 [23] to extract local features, while the auxiliary encoder replaces Transformer with VMamba to capture global semantic features. It strikes a balance between accurate long-range dependency modelling and computational efficiency. To facilitate cross-modal interaction between the two encoders and to exploit complementary information across modalities, we introduce a multimodal feature fusion block (MFFB) at each feature extraction stage. This module integrates global and local information using multi-kernel convolution and window-based cross-attention, ensuring effective fusion of multimodal features.

In the decoder stage, to better integrate deep and shallow features both semantically and in the frequency domain, We employ FreqFusion [24], a frequency-aware guided upsampling method, as a replacement for conventional feature fusion techniques. This approach enhances the restoration of fine spatial details and contributes to improved segmentation accuracy. In summary, the main contributions of this study are as follows:

A novel MFFB is designed to extract complementary features across modalities. The local branch captures fine-grained details using multi-scale depthwise separable convolutions, while the global branch employs Efficient Additive Attention [25] to model global contextual information. These are further integrated via window-based cross-attention to enable effective interaction between local and global representations.

A frequency-aware guided upsampling method (FreqFusion) is introduced in the decoder stage to replace traditional upsampling strategies. This facilitates semantic fusion of deep and shallow features and enables finer reconstruction of spatial details.

Extensive experiments conducted on the ISPRS Vaihingen and Potsdam datasets demonstrate that MMFNet achieves superior segmentation performance compared to eight state-of-the-art methods, while maintaining low computational complexity.

2. Related Work

2.1. Single Modal Semantic Segmentation

In recent years, deep learning has emerged as the dominant approach for semantic segmentation. Fully Convolutional Network (FCN) [26] marked a major milestone as the first end-to-end architecture for pixel-level prediction, significantly improving segmentation accuracy. However, FCNs rely solely on deep semantic features and perform a single-step upsampling, which often leads to blurred object boundaries and limited segmentation precision. To address these limitations, U-Net [27] introduced an encoder–decoder architecture with skip connections. This design enables the progressive fusion of deep semantic features and shallow spatial details, effectively refining object boundaries and enhancing segmentation accuracy.

Many researchers have enhanced the global context modelling capabilities of CNNs by stacking multiple convolutional layers, employing dilated convolutions [28], or introducing attention mechanisms [29]. However, dilated convolutions often fail to capture fine-grained details, thereby weakening the network's ability to segment small objects. Attention mechanisms, while improving contextual awareness, still rely on convolutional backbones and thus remain inherently limited to local information exchange. Since convolution operations extract features within fixed receptive fields, CNN-based methods are fundamentally ineffective in capturing global semantic context and

 doi:10.20944/preprints202508.0078.v1

long-range dependencies [30]. Due to their powerful ability to model long-range visual dependencies, Transformers have attracted increasing attention in computer vision tasks. ViT, divides input images into patches and processes them using a pure Transformer architecture for image classification, achieving promising results. Compared to CNNs, ViT significantly improves the modelling of global contextual information, which has inspired a series of subsequent works [31–34].

However, the quadratic computational complexity of Transformers with respect to input size poses significant challenges when applied to dense prediction tasks such as object detection and semantic segmentation. This has created a pressing need for a novel vision backbone that offers global receptive fields and dynamic weighting while retaining linear complexity.

The recently proposed Mamba model has emerged as a promising alternative to Transformers, owing to its ability to model long-range dependencies with linear computational cost. Recent works such as Vim and VMamba have successfully extended Mamba to visual tasks, demonstrating its effectiveness in image representation learning.

In the context of RSI segmentation, Zhu et al. replaced the multi-head self-attention in ViT with Mamba to capture global dependencies from image data, and integrated it with multilayer perceptron (MLP) to construct the Samba block [35]. Additionally, RS-Mamba [36] introduces a VSS module equipped with an Omnidirectional Selective Scan Module (OSSM), which enhances multi-directional global context modelling and enables comprehensive spatial feature extraction.

2.2. Multimodel Semantic Segmentation (MSS)

The Materials and Methods should be described with sufficient details to allow others to replicate and build on the published results. Please note that the publication of your manuscript implicates that you must make all materials, data, computer code, and protocols associated with the publication available to readers. Please disclose at the submission stage any restrictions on the availability of materials or information. New methods and protocols should be described in detail while well-established methods can be briefly described and appropriately cited.

With advances in Earth observation technologies, the acquisition of multimodal remote sensing data has significantly improved. Increasing attention has been given to the integration of data such as optical imagery, multispectral images, and DSMs. DSMs capture surface height information and serve as key indicators for identifying objects with similar spectral features but different elevations, such as trees and low-lying vegetation in confined areas. The high spatial precision of DSMs enhances boundary delineation, thereby contributing to improved segmentation performance in RSI.

Multimodal fusion strategies are generally classified into three categories: early fusion, intermediate fusion, and late fusion.

Early fusion involves stacking multiple modalities along the channel dimension before feeding them into the network. For instance, ResUNet-a [37] concatenates RGB images (three channels) and DSM data (one channel) into a four-channel input. However, this straightforward channel-wise concatenation fails to fully exploit the complementary nature of multimodal information, often introducing noise or redundancy. As a result, cross-modal feature extraction remains suboptimal, and segmentation accuracy gains are limited.

Late fusion, on the other hand, extracts features from each modality using separate branches and combines their outputs at the decision level—typically via weighted averaging, voting, or ensemble classification. For example, V-FuseNet [38] utilizes two independent branches to process RGB and DSM data separately and fuses them by element-wise addition at the output layer. While this approach preserves modality-specific features, the lack of interaction during feature extraction leads to a loss of cross-modal correlations, limiting the effectiveness of the final prediction.

To overcome the limitations of early and late fusion, most recent studies favor intermediate fusion strategies. This approach typically employs parallel dual-branch networks to extract features from each modality independently, followed by hierarchical fusion at intermediate layers. The fused features are then upsampled using a decoder for final prediction.

 doi:10.20944/preprints202508.0078.v1

FuseNet [39] uses two CNNs to extract features from RGB and depth images, respectively, and integrates depth features into the RGB stream at various levels using a sparse fusion strategy. CMGFNet [11] introduces a GFM to adaptively combine multimodal features while reducing redundancy during fusion. CMFNet [12] leverages a cross-attention mechanism to integrate multimodal information at multiple scales. EDFT [40], built upon SegFormer [32], adopts a two-branch architecture and proposes a depth-aware self-attention (DSA) module to fuse multimodal features. Intermediate fusion not only enables effective intra-modality feature extraction but also progressively aligns features across modalities by integrating information at multiple scales, thereby enhancing the overall representational power of multimodal data.

Mamba combines the ability to model long-range dependencies with the advantage of linear computational complexity, making it a compelling alternative to conventional Transformer-based architectures. Recent studies have explored its potential in multimodal learning tasks. AlignMamba [41] introduces a novel fusion framework that employs local alignment via optimal transport and global alignment via maximum mean discrepancy (MMD) to address the cross-modal misalignment challenge inherent in Mamba's sequential processing. This enhances the efficiency of multimodal feature fusion.

Sigma [20] is the first to apply Mamba to MSS, proposing a Mamba-based cross-attention mechanism to facilitate interaction across modalities. In the context of multimodal RSI segmentation, MFMamba [21] represents a pioneering effort. It utilizes CNNs to extract local features and Mamba to capture global context. However, its fusion strategy—element-wise addition of local and global features—limits the extent of cross-modal interaction and weakens the model's ability to represent complex spatial relationships.

3. Methodology

3.1. Framework of MMFNet

As illustrated in Figure 2, the overall architecture of MMFNet follows the design of MFMamba and consists of four main components: a primary encoder, an auxiliary encoder, a feature fusion module, and a decoder. The primary encoder adopts ResNet-18, which focuses on capturing fine-grained local details, while the auxiliary encoder employs VMamba to provide complementary global context to the features extracted by ResNet-18.

 doi:10.20944/preprints202508.0078.v1

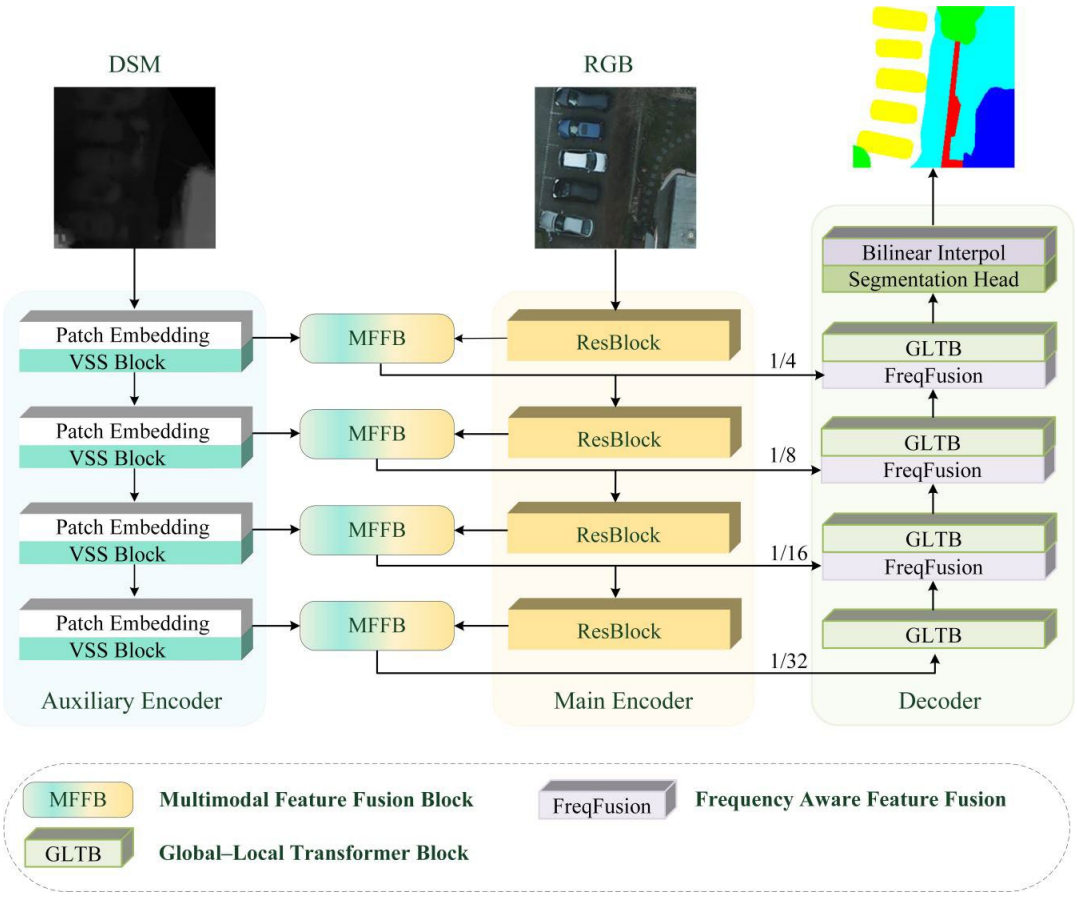


Figure 2. Overall architecture of MMFNet.

The feature fusion module integrates multimodal features at each stage of the encoding process to generate multi-scale, multimodal representations. In MFMamba, feature fusion is implemented using the FFB. However, its element-wise addition strategy may cause information loss or feature misalignment when modality-specific characteristics differ significantly.

To address this limitation, we propose a novel MFFB, which incorporates both global and local branches to capture complementary information across modalities. A window-based cross-attention mechanism is employed within the fusion process to enable rich cross-modal interaction. This strategy facilitates the modelling of both intra- and inter-modal dependencies, allowing for more effective multimodal feature integration.

The decoder is responsible for restoring spatial details from the high-level features generated by the encoder, enabling precise segmentation. While MFMamba adopts bilinear upsampling, its fixed interpolation weights are not adaptive to varying semantic regions, leading to blurred boundaries in cross-scale feature fusion. To overcome this, we replace bilinear interpolation with FreqFusion in the decoder stage. FreqFusion performs frequency-domain decomposition to achieve adaptive fusion between deep and shallow features, thereby improving boundary sharpness in segmentation results.

3.2. Dual Branch Encoder

As shown in Figure 2, MMFNet adopts a dual-encoder architecture that facilitates multimodal feature extraction through the coordinated operation of a primary encoder based on ResNet-18 and an auxiliary encoder based on VMamba. The ResNet-18 encoder focuses on capturing local detail features, while the VMamba encoder—built upon the SSM—models' global contextual information. Together, they form a complementary encoding framework that enables the effective representation of both fine-grained and long-range semantic features.

 doi:10.20944/preprints202508.0078.v1

The primary encoder consists of four stages of residual blocks (ResBlocks). At each stage, features are enhanced by MFFB and then forwarded to the decoder. The auxiliary encoder is constructed using the Visual State Space (VSS) module (as shown in Figure 3), whose core component is the Selective Scanning 2D (SS2D) mechanism. SS2D enables efficient long-range dependency modelling through multi-directional sequential scanning and state space transformation.

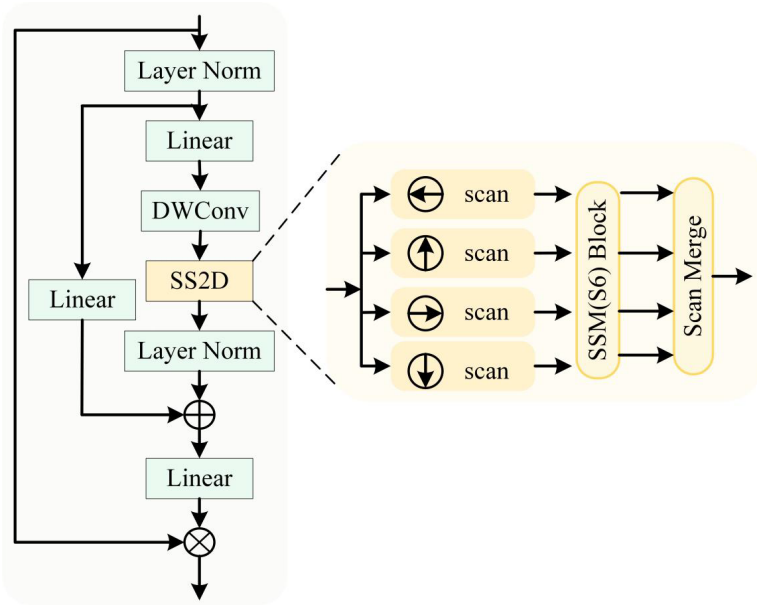


Figure 3. Overall structure of the Visual State Space (VSS) module.

The auxiliary encoder also comprises four stages. It performs hierarchical downsampling through patch embedding and patch merging operations, while the integrated VSS modules progressively learn global semantic representations across scales.

This architectural design integrates the strengths of CNNs in local feature extraction with the global semantic modelling capabilities of Mamba, enabling an effective mechanism for collaborative representation of multimodal features.

3.3. Multimodel Feature Fusion Block

Fusion methods based on simple concatenation or alignment often fail to fully exploit the complementary characteristics of cross modal data [42] while also neglecting the long-range dependencies between multimodal inputs [43]. Moreover, inherent noise and redundant features may further degrade the quality of feature representation [10].

To effectively model both intra-modal and inter-modal dependencies, and to extract modality-specific as well as modality-shared features, we design the MFFB. As illustrated in Figure 4, MFFB consists of three main components: a local branch, a global branch, and a window cross attention (WCA) module. The local and global branches are responsible for capturing the detailed or global features that may be lacking in certain modalities, while the WCA module models cross-modal dependencies to obtain complementary features across modalities.

 doi:10.20944/preprints202508.0078.v1

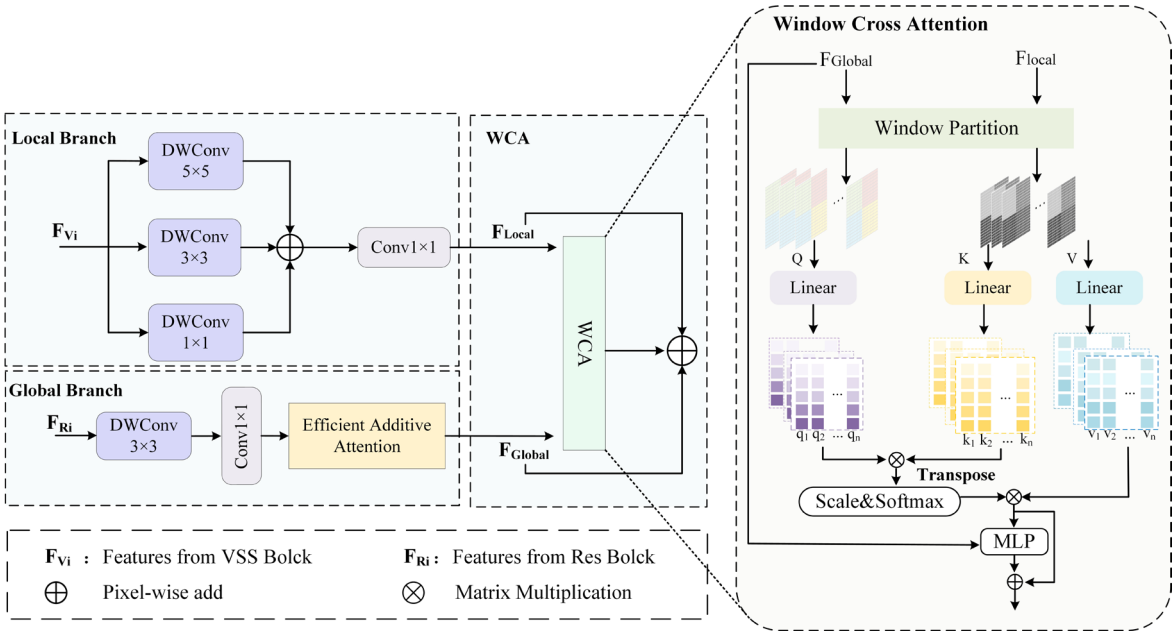


Figure 4. Overall architecture of the Multimodal Feature Fusion Block (MFFB).

The MFFB takes as input the feature maps F_{Ri} from the CNN-based primary encoder and F_{Vi} from the VSS-based auxiliary encoder. Due to the limited receptive field of CNNs, the feature maps F_{Ri} from the primary encoder lack sufficient global context. Conversely, although Mamba excels at modelling long range dependencies, it is less effective in capturing fine-grained local details, resulting in relatively coarse representations in F_{Vi} from the auxiliary encoder. To integrate the global and local features from F_{Ri} and F_{Vi}, MFFB processes F_{Ri} through the global branch, where Efficient Additive Attention (for details, see [25]) is applied to extract global contextual features F_{Global}.

Simultaneously, F_{Vi} is passed through the local branch, where three convolutional layers with different kernel sizes are used to capture fine-grained local features F_{Local}. The outputs of the global and local branches are then fused via the WCA module, as illustrated in Figure 4.

To avoid the high computational cost of global attention, both F_{Local} and F_{Global} are first partitioned into non-overlapping windows of size 7×7. These windowed features are then passed through linear layers to generate the corresponding queries (Q), keys (K), and values (V), which are subsequently used to compute the cross attention. This process can be expressed by Equation (1), (2) and (3) as:

$$Q = W_Q F_{Global}, K = W_K F_{Local}, V = W_V F_{Local} \tag{1}$$

$$Attention(Q,K,V) = \mathrm{SoftMax}(\frac{QK^T}{\sqrt{d}} + B)V \tag{2}$$

$$F_{out} = MLP(Attention(Q,K,V)) \tag{3}$$

here, $B \in R^{W^2 \times W^2}$ denotes the relative positional bias, while $Q,K,V \in R^{W^2 \times d}$ represent the query, key, and value matrices, respectively; d refers to the dimensionality of the query/key vectors, and W is the window size.

The WCA model performs cross attention computation within each local window. By modelling cross modal dependencies in a windowed manner, WCA effectively captures complementary features across modalities with reduced computational overhead.

 doi:10.20944/preprints202508.0078.v1

3.4. Transfomer Decoder

HRRSI poses significant challenges for object recognition and localization due to its high spatial resolution, dense object distribution, and wide variations in object scale. Therefore, the combination of global context and fine spatial detail is critical for reliable semantic reasoning in HRRSI [43]. Although encoder–decoder architectures such as U-Net integrate shallow high-resolution features with deep semantic representations via skip connections, their decoders typically rely on fixed interpolation kernels—such as Bilinear Interpolation or Nearest Neighbor Interpolation—which are insufficient for capturing the rich contextual information required for accurate semantic reasoning. To address this limitation, FreqFusion [24] introduces an adaptive frequency-domain upsampling strategy. Specifically, it applies adaptive low-pass filtering to upsampling deep (low-resolution) features by suppressing high-frequency noise and maintaining semantic consistency. A displacement generator is further used to perform spatial alignment. Meanwhile, adaptive high-pass filtering is applied to enhance boundary details in shallow (high-resolution) features, compensating for the loss of high-frequency information during downsampling. This enables complementary fusion of deep and shallow features in the frequency domain.

The decoder in this study adopts a hybrid architecture combining global local transformer block (GLTB) and FreqFusion. GLTB employs a multi-scale attention mechanism to simultaneously capture long-range dependencies and local spatial details. FreqFusion performs frequency-aware upsampling on deep (low-resolution) features and utilizes an adaptive frequency filtering mechanism in the spatial domain to achieve complementary fusion of deep and shallow features in terms of both semantic consistency and fine-grained detail. Further details of GLTB and FreqFusion can be found in [24,34].

3.5. Loss Function

We employ the cross-entropy loss to supervise the training of the network, which is defined as follows:

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i) \tag{4}$$

where t_i denotes the GT, and p_i is the softmax probability of class i.

4. Experiment and Results

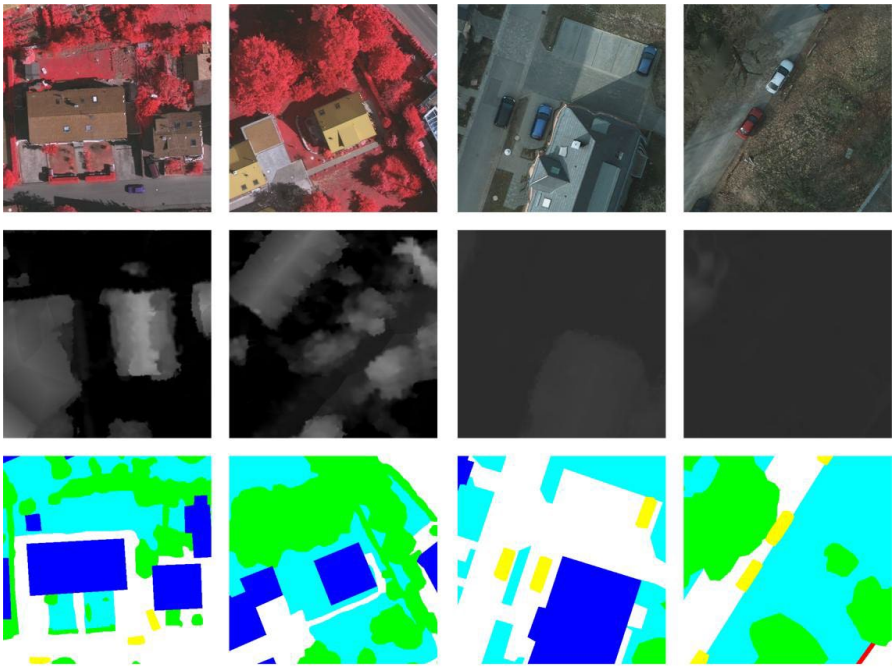
4.1. Dataset

1) Vaihingen: Vaihingen dataset consists of 16 high-resolution true orthophoto images, each with an average size of approximately 2500 × 2000 pixels. Each image contains three spectral channels—near-infrared (NIR), red, and green (NIRRG)—as well as a DSM with a ground sampling distance (GSD) of 9 cm. The dataset includes five foreground classes: building (Bui.), tree (Tre.), low vegetation (Low.), car, and impervious surface (Imp.), as well as one background class: clutter. In our experiments, we utilized TOP image tiles and complete images. The 16 images are divided into a training set of 12 images and a test set of 4 images. The training set includes image IDs: 1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34, and 37; the test set comprises images 5, 21, 15, and 30.

2) Potsdam: Potsdam dataset consists of 24 high-resolution aerial images captured over the city of Potsdam, Germany, each with a resolution of 6000 × 6000 pixels and a GSD of 5 cm. It provides four multispectral channels, including infrared, red, green, and blue (IRRGB), along with a DSM at the same 5 cm GSD. The dataset shares the same semantic classes as the Vaihingen dataset. In our experiments, we use the RGB composite images together with the corresponding DSM data. The dataset is divided into 18 images for training and 16 images for validation or testing. The training set comprises the following image IDs: 6_10, 7_10, 2_12, 3_11, 2_10, 7_8, 5_10, 3_12, 5_12, 7_11, 7_9, 6_9,

 doi:10.20944/preprints202508.0078.v1

7_7, 4_12, 6_8, 6_12, 6_7, and 4_11. The test/validation set includes: 2_11, 3_10, 6_10, 7_10, 2_12, and 3_11. Figure 5 presents some data samples from the Vaihingen and Potsdam datasets.



(**a**) (**b**) (**c**) (**d**)

Figure 5. (**a**), (**b**) are sample images from the Vaihingen dataset; (**c**), (**d**) are from the Potsdam dataset. The first row shows the orthophotos with three spectral channels (NIRG for Vaihingen and RGB for Potsdam). The second and third rows present the corresponding DSM data and pixel-wise semantic labels, respectively.

4.2. Evaluation Metrics

To quantitatively assess the performance of segmentation, we adopt four commonly used metrics: Intersection over Union (IoU), mean IoU (mIoU), overall accuracy (OA), and mean F1-score (mF1). These metrics are computed based on four fundamental quantities: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

For each class, IoU is defined as the ratio between the intersection and the union of the predicted and GT regions, and is calculated as follows:

$$IOU = \frac{TP}{TP + FP + FN} \tag{5}$$

The F1-score for each class is calculated as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

The precision for each class is calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

The recall for each class is calculated as follows:

$$recall = \frac{TP}{TP + FN} \tag{8}$$

In addition, mIoU refers to the average IoU across all classes, and mF1-score denotes the mF1 calculated over all categories.

 doi:10.20944/preprints202508.0078.v1

4.3. Experiment Setup

All experiments were implemented using PyTorch and conducted on a single RTX 4080 GPU. During training, images were randomly cropped into 256 × 256 patches, and data augmentation techniques such as random horizontal flipping, random vertical flipping, and random rotation were applied. The number of training epochs was set to 50.

The model was optimized using Stochastic Gradient Descent (SGD) with a learning rate of 0.01, a momentum of 0.9, a weight decay of 0.0005, and a batch size of 16.

4.4. Experimental Results and Analysis

4.4.1. Comparison Results on the Vaihingen Dataset

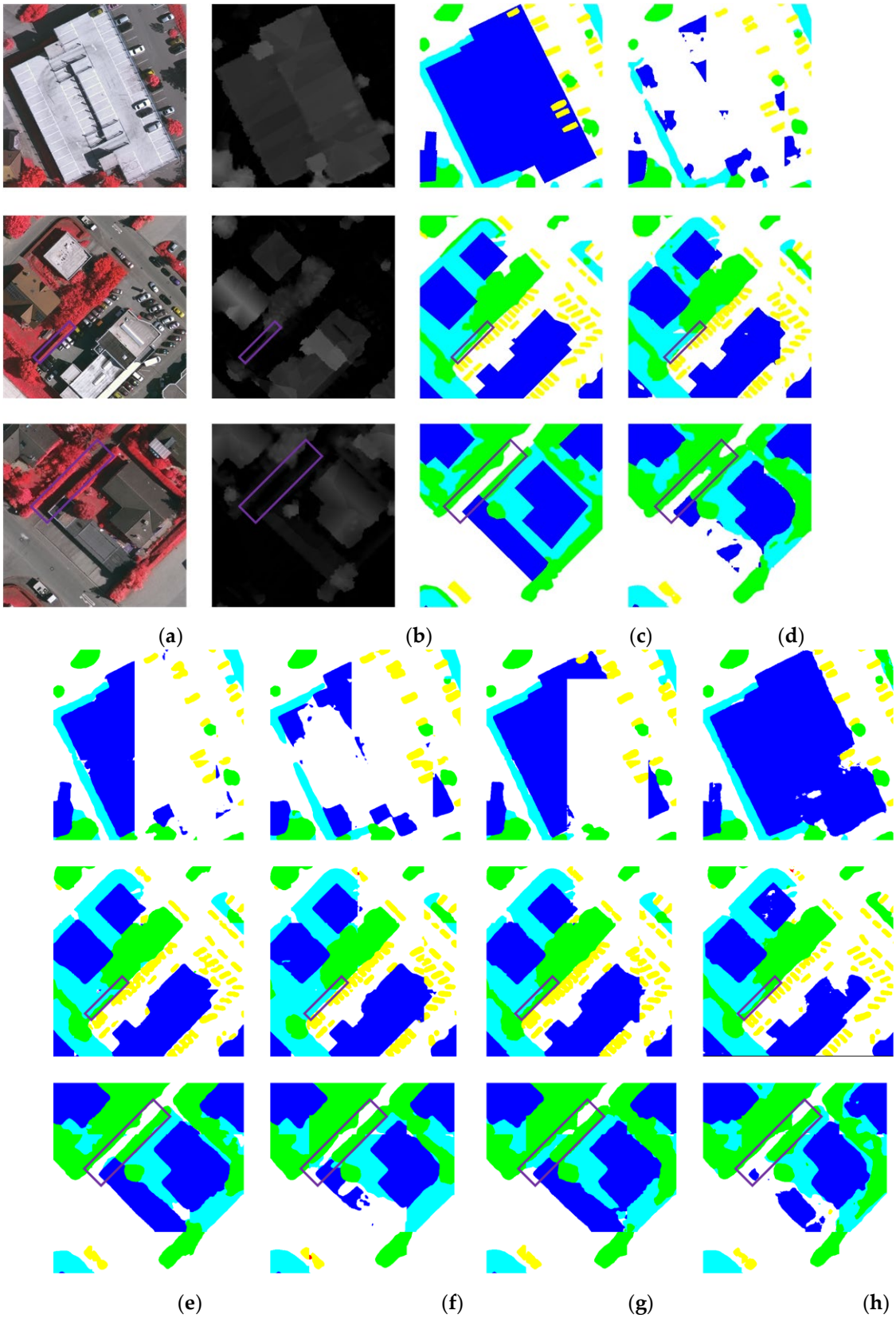
As shown in Table 1, the proposed MMFNet achieves the highest scores on the Vaihingen dataset in terms of OA, mF1, and mIoU. Compared with the baseline MFMamba, MMFNet shows consistent improvements of 1.59% in mF1 and 2.04% in mIoU, demonstrating its ability to effectively extract and fuse complementary features from DSM and HRRSI data. In comparison with existing SOTA methods, MMFNet also delivers superior segmentation performance in key categories such as building, tree, low vegetation, and impervious surface. Specifically, it achieves an increase of 2.41% in IoU for low vegetation and 1.21% in IoU for car over the baseline, further highlighting its advantage in modelling fine-grained multimodal features.

Table 1. Comparison results with other methods on the Vaihingen dataset. Bold numbers indicate the optimal value, and underlined numbers represent the sub-optimal value. (Unit: %).

Model	Backbone	Imp.	Bui.	Low.	Tre.	Car	OA	mF1	mIoU
		IoU							
PSPNet	Resnet-18	79.27	89.5	60.41	77.25	71.45	87.58	86.31	75.76
Swin	Swin-T	81.49	89.93	63.08	75.05	64.97	86.74	84.87	74.90
Unetformer	Resnet-18	79.33	88.67	61.86	73.56	70.73	86.65	85.31	74.83
DCSwin	Swin-T	81.47	89.81	63.69	74.82	70.54	87.70	86.11	76.07
CMFNet	VGG-16	86.59	94.25	66.75	82.75	77.03	91.39	89.50	81.47
Vmamba	Vmamba-T	82.58	89.06	67.35	77.83	57.07	82.23	84.56	74.78
RS3Mamba	R18-Mamba-T	85.21	92.34	66.51	82.94	81.21	90.87	89.28	81.64
MFMamba	R18-Mamba-T	86.04	94.02	66.14	83.64	77.43	91.37	89.48	81.46
Ours	R18-Mamba-T	87.55	94.37	68.92	84.21	82.42	92.06	90.77	83.50

Figure 6 presents visual comparisons of the segmentation results obtained by the nine methods. In the first and second rows, several single-modality models misclassify regions within buildings as impervious surfaces. In contrast, our method and CMFNet yield the most accurate building predictions. In the second row, the building contours predicted by our model are more consistent with the GT, exhibiting smoother and more complete edges. This demonstrates the ability of MMFNet to more precisely segment large-scale structures. Additionally, within the dashed box in the second row, MMFNet successfully identifies trees located within areas of low vegetation, whereas other methods (Figure 6 (**d**)–(**k**)) either miss or partially detect them. This result highlights MMFNet's superior effectiveness in addressing the challenge of inter-class similarity. In the third row, the dashed box marks a region where shadows cast by trees on both sides of the road alter the appearance of the surface in RGB images, making the road's color and texture significantly different from the surrounding areas. As a result, most comparison methods misclassify the region. However, MMFNet accurately distinguishes the road from adjacent trees, demonstrating its robustness in mitigating the adverse effects of shadow occlusion.

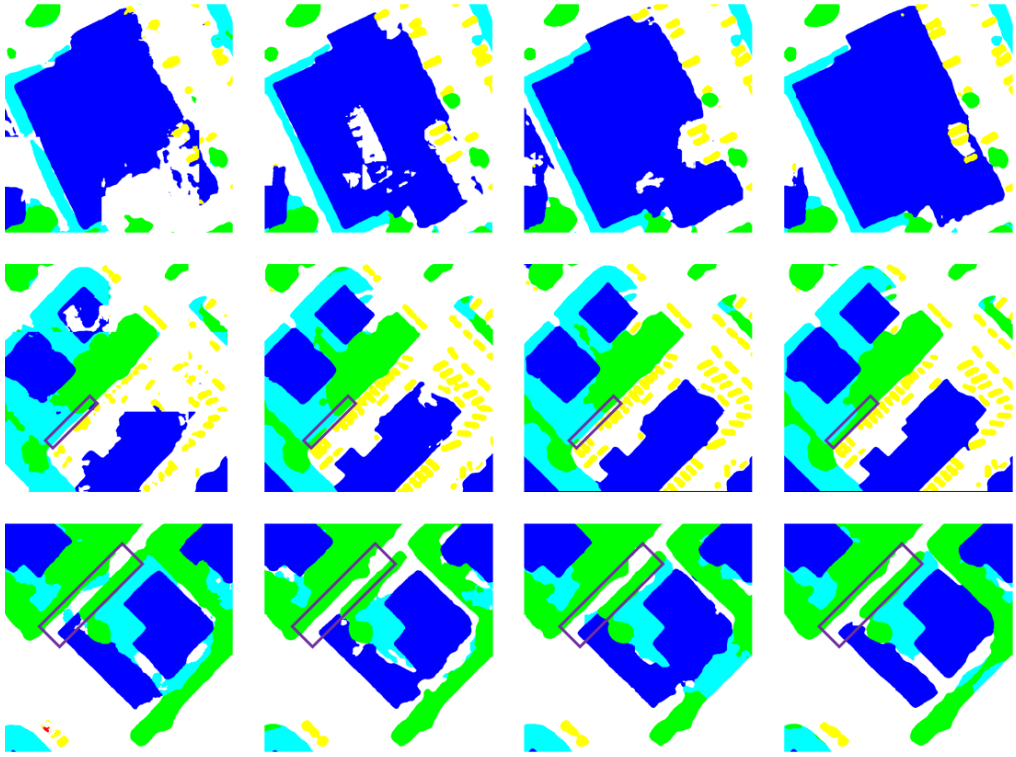
 doi:10.20944/preprints202508.0078.v1



(a) (b) (c) (d)

(e) (f) (g) (h)

 doi:10.20944/preprints202508.0078.v1



(i) (j) (k) (l)

Figure 6. Visual comparison of segmentation results on the Vaihingen dataset using different methods. (a) NIRRG; (b) DSM; (c) GT; (d) PSPNet; (e) Swin; (f) UNetFormer; (g) DCSwin; (h) CMFNet; (i) VMamba; (j) RS3Mamba; (k) MFMamba; (l) Ours.

4.4.2. Comparison Results on the Potsdam Dataset

The experimental results on the Potsdam dataset are consistent with those on the Vaihingen dataset, where our method achieves the highest scores in terms of OA, mF1, and mIoU. Compared with the baseline MFMamba, our method improves OA, mF1, and mIoU by 0.43%, 0.39%, and 0.65%, respectively. Notably, it also demonstrates superior segmentation performance in key categories such as building, tree, low vegetation, and impervious surface when compared with other SOTA methods. Specifically, our method yields an improvement of 1.45% in IoU for low vegetation and 0.81% in IoU for buildings over the baseline.

Table 2. Comparison results with other methods on the Potsdam dataset (Unit: %). Bold numbers indicate the optimal value, and underlined numbers represent the sub-optimal value.

Model	Backbone	IoU					OA	mF1	mIoU
		Imp.	Bui.	Low.	Tre.	Car			
PSPNet	Resnet-18	78.98	88.93	68.23	68.42	77.77	86.12	82.51	76.47
Swin	Swin-T	79.28	90.5	69.98	70.41	79.64	87.05	83.69	77.96
Unetformer	Resnet-18	84.51	92.08	72.70	71.42	83.45	89.19	89.20	80.83
DCSwin	Swin-T	82.96	92.50	71.31	71.24	82.29	88.31	88.71	80.06
CMFNet	VGG-16	85.55	93.65	72.23	74.65	91.25	89.97	91.01	83.37
Vmamba	Vmamba-T	84.82	91.24	75.16	75.38	88.04	81.52	90.40	82.93
RS3Mamba	R18-Mamba-T	86.95	94.46	75.50	76.28	92.98	90.73	87.39	85.24
MFMamba	R18-Mamba-T	87.34	94.90	75.09	76.81	92.88	90.89	91.92	85.41
Ours	R18-Mamba-T	87.41	95.71	76.54	77.50	93.13	91.32	92.31	86.06

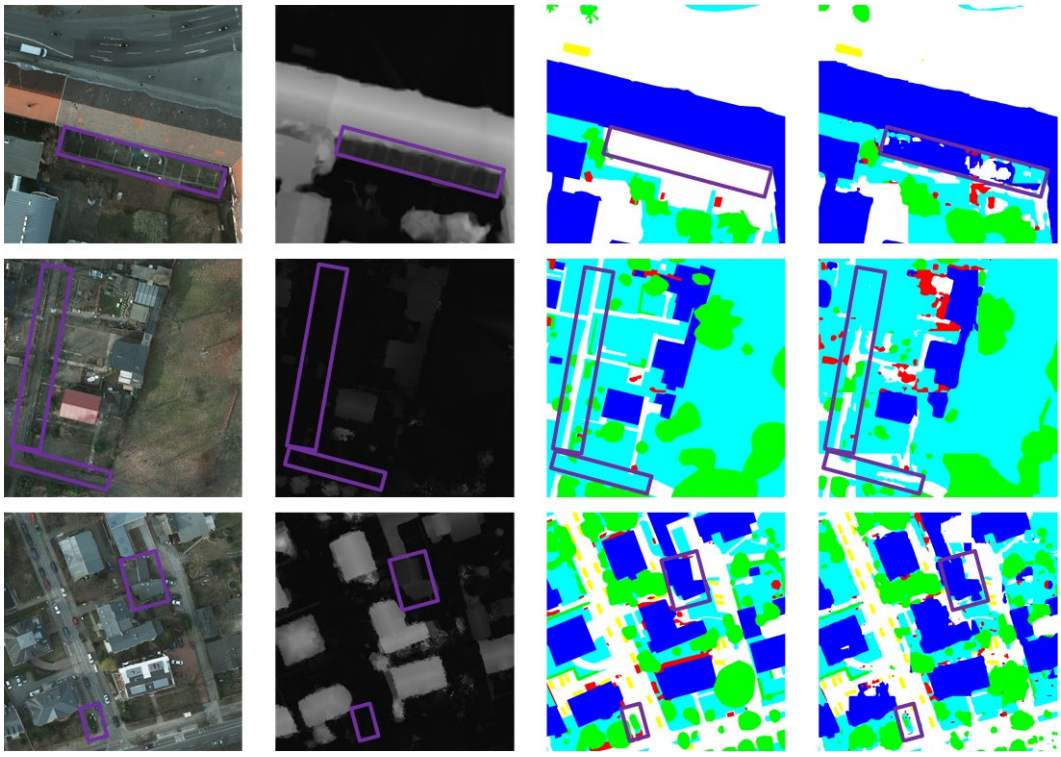
Figure 7 presents visual examples of the segmentation results produced by all nine methods on the Potsdam dataset. Although the performance differences in terms of OA and mIoU are not

 doi:10.20944/preprints202508.0078.v1

particularly large, our method shows noticeably better performance in segmenting large buildings and visually confusing regions compared to other networks. As shown in the top-right rectangular windows of the first and third rows, MMFNet is able to accurately identify the entire building structure, whereas other methods produce fragmented results or mistakenly classify other objects as buildings. From the rectangular boxes in the second and third rows, it is evident that our method performs better in confusing areas than the others. In the second row, the road highlighted in the rectangular window is narrow and surrounded by dense vegetation, making it easy to be confused with the surrounding vegetation. MMFNet successfully distinguishes the road from the vegetation, while other methods tend to confuse the two. Low vegetation and trees are also commonly confused due to their inter-class similarity. In the bottom-left rectangular box of the third row, where trees are surrounded by low vegetation, MMFNet is able to accurately separate the two categories, whereas other methods tend to confuse them.

4.4.3. Computational Complexity Analysis

We adopt floating point operations (FLOPs) and the number of model parameters as evaluation metrics to assess the computational complexity of the proposed MMFNet. FLOPs serve as an indicator of the time complexity of deep learning-based models, while the parameter count quantifies model size. Table 3 presents the complexity analysis results for all comparison methods in this study.



(**a**) (**b**) (**c**) (**d**)

 doi:10.20944/preprints202508.0078.v1

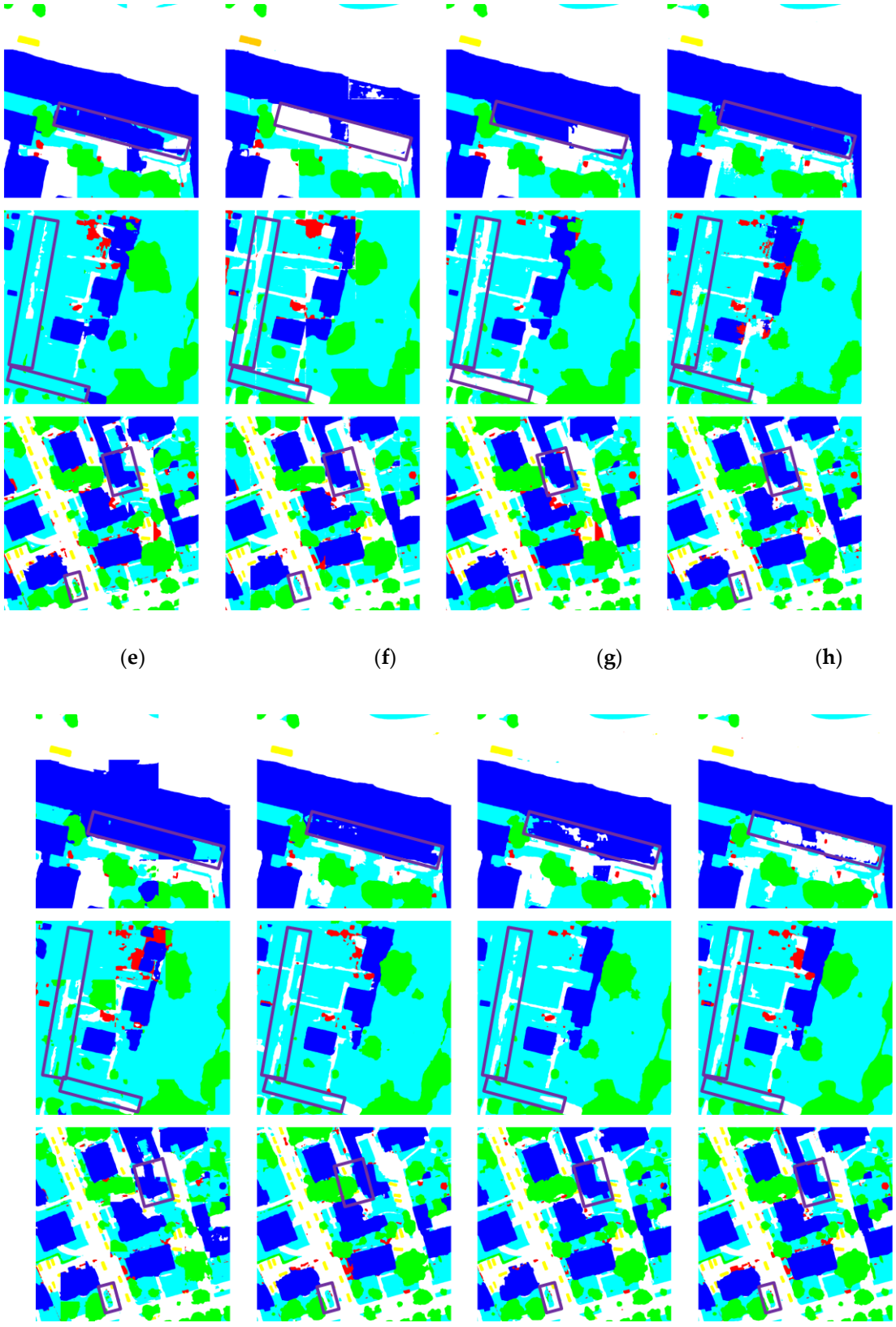


Figure 7. Visual comparison of segmentation results on the Potsdam dataset using different methods. (**a**)RGB; (**b**)DSM; (**c**)GT; (**d**)PSPNet; (**e**)Swin; (**f**)UNetFormer; (**g**)DCSwin; (**h**)CMFNet ; (**i**)VMamba; (**j**) RS3Mamba; (**k**) MFMamba; (**l**)Ours.

 doi:10.20944/preprints202508.0078.v1

Table 3. Comparison of Computational Complexity and mIoU Performance on the Vaihingen Dataset.

Method	FLOPs(G)	Parameter(M)	mIoU(%)
PSPNet	64.15	65.60	75.76
Swin	60.28	59.02	74.90
Unetformer	5.99	11.72	74.83
DCSwin	40.08	66.95	76.07
CMFNet	159.55	104.07	81.47
Vmamba	12.41	29.94	74.78
RS3Mamba	19.78	43.32	81.64
MFMamba	19.12	62.43	81.46
Ours	19.15	69.85	83.50

As shown in Table 3, although UnetFormer has the lowest FLOPs and parameter count, its mIoU score is significantly lower than that of our model. Compared with Transformer-based multimodal methods such as CMFNet, MMFNet achieves a substantial reduction in FLOPs and requires fewer parameters while maintaining superior mIoU performance. This efficiency is primarily attributed to the use of Mamba as the auxiliary branch in the encoder, in contrast to Transformer, which is typically more resource-intensive.

When compared with MFMamba, another multimodal method based on RS3Mamba, MMFNet has a larger number of parameters, but the FLOPs remain the same, and the segmentation performance is significantly improved. Compared to single-modality segmentation methods, the computational complexity of our model is slightly higher due to the incorporation of multimodal data, yet it achieves notably better segmentation performance. Furthermore, relative to the baseline RS^{3}Mamba, MMFNet yields substantial improvements in segmentation accuracy with only a modest increase in the number of model parameters.

4.5. Ablation Studies

To evaluate the effectiveness of incorporating DSM data, we conduct ablation experiments by setting the input to HRRSI only and HRRSI+DSM on Vaihingen and Potsdam datasets. As shown in Table 4, the inclusion of DSM data leads to overall performance improvements of MMFNet on both datasets. In particular, the most significant gains are observed in the segmentation of impervious surfaces and buildings, which can be attributed to the stable elevation characteristics of these classes.

Table 4. Ablation Study on the Effect of DSM Data on the Vaihingen and Potsdam Datasets.

Dataset	Bands	Class OA (%)					mF1(%)	mIoU(%)
		Imp.	Bui.	Low.	Tre.	Car		
Vaihingen	NIRRG	92.04	96.06	80.64	92.04	88.36	90.04	82.29
	NIRRG+DSM	93.57	97.21	81.16	91.50	88.39	90.77	83.50
Potsdam	RGB	92.45	97.50	88.94	86.63	96.11	91.63	84.91
	RGB+DSM	92.70	98.01	88.78	87.65	96.38	92.31	86.06

However, the segmentation accuracy for low vegetation and trees shows slightly different trends across the two datasets. Specifically, the inclusion of DSM improves the performance of MMFNet on low vegetation in the Vaihingen dataset but reduces its accuracy on trees. Conversely, in the Potsdam dataset, adding DSM decreases the performance on low vegetation but enhances the segmentation of trees. This may be due to the strong spectral similarity between these two classes, along with their highly irregular structures and ambiguous boundary shapes. These results suggest that while DSM contributes to improved overall segmentation performance, distinguishing between trees and low vegetation remains a challenge due to their inherent inter-class similarity and spatial complexity.

 doi:10.20944/preprints202508.0078.v1

To validate the effectiveness of the proposed WCA module and the integration of FreqFusion, we conduct ablation experiments by comparing the model performance with different components added. The evaluation results are summarized in Table 5, where a tick (√) indicates that the corresponding module is included. The first row of Table 5 presents the ablation result for FreqFusion, where bilinear interpolation is used to upsample feature maps instead of the proposed frequency-aware method. The second row shows the ablation result for WCA. In this setting, the structure of MFFB is retained, but the fusion operation is modified: the local feature F_{Local} and the global feature F_{Global} are fused via element-wise addition without the WCA module.

Table 5. Ablation Study of WCA and FreqFusion on the Potsdam Dataset (Unit: %).

WCA	FreqFusion	Imp.	Bui.	Low.	Tre.	Car	OA	mF1	mIoU
√	×	87.53	95.49	75.71	77.48	93.30	91.13	92.21	85.90
×	√	87.32	95.31	75.44	77.50	93.09	90.97	92.12	85.73
√	√	87.41	95.71	76.54	77.50	93.13	91.32	92.31	86.06

The results in the first and third rows demonstrate that FreqFusion enables effective complementary fusion between deep and shallow features. The notable improvements in global semantics (e.g., building) and local details (e.g., low vegetation) suggest that this module enhances the model's capability to represent multi-scale features.

The results in the second and third rows confirm that the WCA module improves semantic understanding in complex scenes by adaptively fusing depth and RGB features. The performance gains for confusing categories such as low vegetation indicate that WCA enhances the quality of cross-modal feature interaction.

Overall, the third-row results show that both components contribute positively to model accuracy, and their combination yields the best performance.

The results in the first and third rows demonstrate that FreqFusion enables complementary fusion between deep and shallow features. The significant improvements in global semantics (building) and local details (low vegetation) indicate that this module enhances the model's ability to represent multi-scale features. The results in the second and third rows verify that the WCA module improves semantic understanding in complex scenes by adaptively fusing DSM and HRRSI features. The observed improvements in confusing categories such as low vegetation suggest that WCA enhances the quality of cross-modal feature interaction. Overall, the third row confirms that both components contribute positively to model accuracy, and their combination yields the best performance.

5. Conclusions

In this study, we proposed MMFNet, a Mamba-based MSS network for remote sensing, designed to address the challenges of complex scene understanding, multimodal feature fusion, and computational efficiency. By integrating the strengths of CNN and Mamba architectures, MMFNet effectively leverages high-resolution spectral information (HRRSI) and DSM data to improve segmentation accuracy while maintaining low computational complexity. To better fuse multimodal features, we introduced the MFFB, which employs a window-based cross-attention mechanism to achieve adaptive fusion across modalities. This design effectively alleviates the insufficient interaction between global and local information in traditional approaches. Additionally, a frequency-aware upsampling module is incorporated into the decoder to reduce the loss of spatial detail during conventional upsampling and to facilitate semantic fusion of deep and shallow features, thereby enhancing edge segmentation accuracy. Compared to existing CNN or Transformer based methods, MMFNet offers a novel perspective for semantic segmentation of multimodal RSI. Experimental results on two public benchmarks ISPRS Vaihingen and Potsdam demonstrate that MMFNet outperforms eight state-of-the-art methods in segmentation performance, while maintaining relatively low computational cost.

 doi:10.20944/preprints202508.0078.v1

Nonetheless, due to differences in imaging mechanisms, feature misalignment between DSM and RGB (IRRG) data can lead to misclassification in the segmentation results. In future work, we plan to explore cross-modal feature alignment techniques to more effectively exploit the complementary information across modalities and further improve the segmentation accuracy of HSRSI.

Author Contributions: Conceptualization, J.Q. and R.Y.; methodology, J.Q. and R.Y.; software, J.Q.; validation, W.C., W.R. and S.H.; formal analysis, W.C.; investigation, S.H.; resources, W.C.; data curation, S.H.; writing—original draft preparation, J.Q.; writing—review and editing, J.Q. and R.Y.; visualization, S.H.; supervision, W.C.; project administration, R.Y.; funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project of Sichuan Bureau of Geology & Mineral Resources, grant number SCDZ-KJXM202514.

Data Availability Statement: The Vaihingen and the Potsdam datasets can be obtained from https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx (accessed on 1 October 2024).

Acknowledgments: We would like to acknowledge the International Society for Photogrammetry and Remote Sensing (ISPRS) for providing the datasets (Vaihingen and Potsdam).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sensing* **2022**, *60*, 1–20. https://doi.org/10.1109/TGRS.2022.3144894.
2. Boonpook, W.; Tan, Y.; Xu, B. Deep Learning-Based Multi-Feature Semantic Segmentation in Building Extraction from Images of UAV Photogrammetry. *International Journal of Remote Sensing* **2021**, *42*, 1–19. https://doi.org/10.1080/01431161.2020.1788742.
3. Weiss, M.; Jacob, F.; Duveiller, G. Remote Sensing for Agricultural Applications: A Meta-Review. *Remote Sensing of Environment* **2020**, *236*, 111402. https://doi.org/10.1016/j.rse.2019.111402.
4. Asadzadeh, S.; Oliveira, W.J.D.; Souza Filho, C.R.D. UAV-Based Remote Sensing for the Petroleum Industry and Environmental Monitoring: State-of-the-Art and Perspectives. *Journal of Petroleum Science and Engineering* **2022**, *208*, 109633. https://doi.org/10.1016/j.petrol.2021.109633.
5. Grekousis, G. Local Fuzzy Geographically Weighted Clustering: A New Method for Geodemographic Segmentation. *International Journal of Geographical Information Science* **2021**, *35*, 152–174. https://doi.org/10.1080/13658816.2020.1808221.
6. Zhou, X.; Zhou, L.; Gong, S.; Zhong, S.; Yan, W.; Huang, Y. Swin Transformer Embedding Dual-Stream for Semantic Segmentation of Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* **2024**, *17*, 175–189. https://doi.org/10.1109/JSTARS.2023.3326967.
7. Jiang, J.; Feng, X.; Ye, Q.; Hu, Z.; Gu, Z.; Huang, H. Semantic Segmentation of Remote Sensing Images Combined with Attention Mechanism and Feature Enhancement U-Net. *International Journal of Remote Sensing* **2023**, *44*, 6219–6232. https://doi.org/10.1080/01431161.2023.2264502.
8. Lin, R.; Zhang, Y.; Zhu, X.; Chen, X. Local-Global Feature Capture and Boundary Information Refinement Swin Transformer Segmentor for Remote Sensing Images. *IEEE Access* **2024**, *12*, 6088–6099. https://doi.org/10.1109/ACCESS.2024.3350645.

 doi:10.20944/preprints202508.0078.v1

9. Qin, R.; Fang, W. A Hierarchical Building Detection Method for Very High Resolution Remotely Sensed Images Combined with DSM Using Graph Cut Optimization. *photogramm eng remote sensing* **2014**, *80*, 873–883. https://doi.org/10.14358/PERS.80.9.873.
10. Cao, Z.; Fu, K.; Lu, X.; Diao, W.; Sun, H.; Yan, M.; Yu, H.; Sun, X. End-to-End DSM Fusion Networks for Semantic Segmentation in High-Resolution Aerial Images. *IEEE Geosci. Remote Sensing Lett.* **2019**, *16*, 1766–1770. https://doi.org/10.1109/LGRS.2019.2907009.
11. Hosseinpour, H.; Samadzadegan, F.; Javan, F.D. CMGFNet: A Deep Cross-Modal Gated Fusion Network for Building Extraction from Very High-Resolution Remote Sensing Images. *ISPRS Journal of Photogrammetry and Remote Sensing* **2022**, *184*, 96–115. https://doi.org/10.1016/j.isprsjprs.2021.12.007.
12. Ma, X.; Zhang, X.; Pun, M.-O. A Crossmodal Multiscale Fusion Network for Semantic Segmentation of Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* **2022**, *15*, 3463–3474. https://doi.org/10.1109/JSTARS.2022.3165005.
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. https://doi.org/10.1145/3065386.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv 2020, arXiv:2010.11929.
15. Lin, G.; Liu, F.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-Path Refinement Networks for Dense Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, 1–1. https://doi.org/10.1109/TPAMI.2019.2893630.
16. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. https://doi.org/10.1145/3505244.
17. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv 2023, arXiv:2312.00752.
18. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. arXiv 2024, arXiv:2401.09417.
19. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual State Space Model. arXiv 2024, arXiv:2401.10166.
20. Wan, Z.; Zhang, P.; Wang, Y.; Yong, S.; Stepputtis, S.; Sycara, K.; Xie, Y. Sigma: Siamese Mamba Network for Multi-Modal Semantic Segmentation 2024. arXiv 2024, arXiv:2404.04256.
21. Wang, Y.; Cao, L.; Deng, H. MFMamba: A Mamba-Based Multi-Modal Fusion Network for Semantic Segmentation of Remote Sensing Images. *Sensors* **2024**, *24*, 7266. https://doi.org/10.3390/s24227266.
22. Cao, Z.; Diao, W.; Sun, X.; Lyu, X.; Yan, M.; Fu, K. C3Net: Cross-Modal Feature Recalibrated, Cross-Scale Semantic Aggregated and Compact Network for Semantic Segmentation of Multi-Modal High-Resolution Aerial Images. *Remote Sensing* **2021**, *13*, 528. https://doi.org/10.3390/rs13030528.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, June 2016; pp. 770–778.
24. Chen, L.; Fu, Y.; Gu, L.; Yan, C.; Harada, T.; Huang, G. Frequency-Aware Feature Fusion for Dense Image Prediction 2024.
25. Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.-H.; Khan, F.S. SwiftFormer: Efficient Additive Attention for Transformer-Based Real-Time Mobile Vision Applications. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Paris, France, October 1 2023; pp. 17379–17390.

 doi:10.20944/preprints202508.0078.v1

26. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sensing Lett.* **2018**, *15*, 474–478. https://doi.org/10.1109/LGRS.2018.2795531.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2015; Vol. 9351, pp. 234–241 ISBN 978-3-319-24573-7.
28. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision – ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2018; Vol. 11211, pp. 833–851 ISBN 978-3-030-01233-5.
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE: Salt Lake City, UT, June 2018; pp. 7132–7141.
30. Wang, D.; Yang, R.; Zhang, Z.; Liu, H.; Tan, J.; Li, S.; Yang, X.; Wang, X.; Tang, K.; Qiao, Y.; et al. P-Swin: Parallel Swin Transformer Multi-Scale Semantic Segmentation Network for Land Cover Classification. *Computers & Geosciences* **2023**, *175*, 105340. https://doi.org/10.1016/j.cageo.2023.105340.
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Montreal, QC, Canada, October 2021; pp. 9992–10002.
32. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Neural Information Processing Systems (NeurIPS)* **2021**.
33. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Montreal, QC, Canada, October 2021; pp. 7242–7252.
34. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **2022**, *190*, 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008.
35. Zhu, Q.; Cai, Y.; Fang, Y.; Yang, Y.; Chen, C.; Fan, L.; Nguyen, A. Samba: Semantic Segmentation of Remotely Sensed Images with State Space Model. *Heliyon* **2024**, *10*, e38495. https://doi.org/10.1016/j.heliyon.2024.e38495.
36. Chi, K.; Guo, S.; Chu, J.; Li, Q.; Wang, Q. RSMamba: Biologically Plausible Retinex-Based Mamba for Remote Sensing Shadow Removal. *IEEE Trans. Geosci. Remote Sensing* **2025**, *63*, 1–10. https://doi.org/10.1109/TGRS.2025.3526966.
37. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS Journal of Photogrammetry and Remote Sensing* **2020**, *162*, 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013.
38. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very High Resolution Urban Remote Sensing with Multimodal Deep Networks. *ISPRS Journal of Photogrammetry and Remote Sensing* **2018**, *140*, 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011.
39. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision – ACCV 2016*; Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2017; Vol. 10111, pp. 213–228 ISBN 978-3-319-54180-8.

 doi:10.20944/preprints202508.0078.v1

40. Yan, L.; Huang, J.; Xie, H.; Wei, P.; Gao, Z. Efficient Depth Fusion Transformer for Aerial Image Semantic Segmentation. *Remote Sensing* **2022**, *14*, 1294. https://doi.org/10.3390/rs14051294.
41. Li, Y.; Xing, Y.; Lan, X.; Li, X.; Chen, H.; Jiang, D. AlignMamba: Enhancing Multimodal Mamba with Local and Global Cross-Modal Alignment. arXiv 2024, arXiv:2412.00833.
42. Chen, Y.; Wang, Q.; Zhao, Y.; Xiong, S.; Lu, X. Bilinear Parallel Fourier Transformer for Multimodal Remote Sensing Classification. *IEEE Trans. Geosci. Remote Sensing* **2025**, *63*, 1–14. https://doi.org/10.1109/TGRS.2025.3556088.
43. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sensing* **2021**, *59*, 426–435. https://doi.org/10.1109/TGRS.2020.2994150.