Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.085484

ARTICLE

Multiscale Long-Distance Feature Aggregation Network for Geospatial Semantic Segmentation in High-Resolution Remote Sensing Imagery

Guangyu Xu[1,2], Yuxi Ban[1], Legend Zhang[3], Junmin Lyu[3], Feng Bao[4] and Wenfeng Zheng[1,3,*]

[1]School of Automation, University of Electronic Science and Technology of China, Chengdu, China
[2]School of the Environment, The University of Queensland, St Lucia, QLD, Australia
[3]Future Tech Institute, Guangzhou Huashang University, Guangzhou, China
[4]School of Biological and Environmental Engineering, Xi'an University, Xi'an, China

*Corresponding Author: Wenfeng Zheng. Email: winfirms@ieee.org or winfirms@uestc.edu.cn

Received: 12 May 2026; Accepted: 25 June 2026; Published: 27 July 2026

ABSTRACT: High-resolution remote sensing semantic segmentation is a fundamental task in Geospatial Artificial Intelligence (GeoAI). Existing CNN-based methods are effective for local and multiscale feature extraction but often lack progressive cross-scale semantic propagation, while attention- and Transformer-based methods improve global spatial modeling but generally ignore frequency-domain regularities. To address these limitations, this study proposes a Multiscale Long-Distance Feature Aggregation Network (MLFANet), a unified spatial-frequency segmentation framework for high-resolution remote sensing imagery. MLFANet introduces three key components: a Multiscale Global Dependency Extraction module for cascaded cross-scale contextual refinement, an FFT-based frequency-domain branch with learnable global filtering for capturing structural and texture regularities, and a bidirectional Spatial-Frequency Fusion module for adaptively aligning spatial details with frequency responses. Experiments on the ISPRS Potsdam and Vaihingen datasets demonstrate the effectiveness and feasibility of the proposed model. MLFANet achieves AF, MIoU, and OA values of 86.03%, 76.21%, and 88.70% on Potsdam, and 83.17%, 71.90%, and 86.33% on Vaihingen, respectively, outperforming representative CNN-based, attention-based, and hybrid models in overall metrics. In terms of computational complexity, MLFANet requires 17.49 G FLOPs under an input size of 256 × 256 pixels, indicating its practical feasibility for patch-based high-resolution remote sensing segmentation. Ablation studies further verify that multiscale dependency extraction, frequency-domain modeling, and adaptive spatial-frequency fusion each contribute to the final performance.

KEYWORDS: Geospatial artificial intelligence (GeoAI); high-resolution remote sensing images; semantic segmentation; spatial-frequency fusion; multiscale feature aggregation; attention mechanism

1 Introduction

The rapid development of geospatial artificial intelligence (GeoAI) has significantly advanced the automatic interpretation of high-resolution remote sensing imagery. As a fundamental task in GeoAI, semantic segmentation aims to assign a land-cover or land-use label to each pixel, thereby supporting a wide range of downstream applications [1–4]. Compared with natural-scene semantic segmentation, geospatial semantic segmentation in high-resolution aerial or satellite imagery involves more complex imaging conditions and spatial organizations. Objects are observed from a nadir or near-nadir perspective, where buildings, roads, trees, vehicles, shadows, and low vegetation are densely interleaved within the same scene. In addition, high-resolution remote sensing images often contain multi-band radiometric variations,

illumination differences, seasonal changes, topological shadows, repeated artificial textures, and severe scale imbalance between large-area objects such as buildings or impervious surfaces and small objects such as cars. These characteristics make high-resolution remote sensing segmentation not merely a general dense prediction problem, but a geospatial interpretation task requiring scale-aware, context-aware, and structure-preserving representation learning [5,6], which is also consistent with recent studies on context-aware aerial landslide extraction and adaptive frequency–spatial interaction for aerial image semantic segmentation [7,8].

Despite remarkable progress in deep learning-based semantic segmentation, high-resolution remote sensing imagery still poses challenges that are not fully addressed by standard computer vision architectures. CNN-based segmentation networks are effective in extracting local textures, object boundaries, and hierarchical semantic features. Representative multiscale methods, including A2-FPN [9], improved PSPNet [10], MCSNet [11], and MAResUNet [12], enhance contextual perception through feature pyramids, pyramid pooling, multi-branch encoding, or attention-enhanced encoder–decoder structures. However, most of these methods rely on parallel fusion, skip connections, or direct concatenation for multiscale aggregation, which may limit progressive semantic interaction across different receptive fields. This limitation is particularly important for high-resolution aerial scenes, where large buildings, roads, vegetation regions, and small vehicles coexist within the same image.

Attention mechanisms and Transformer-based architectures have been introduced to strengthen long-distance dependency modeling. Liu et al. [13] analyzed the applicability of Transformers to remote sensing semantic segmentation, while STUNet [14] and CMTFNet [15] explored Transformer-based or CNN–Transformer hybrid structures for hierarchical contextual modeling. Recent attention-enhanced models, including AFENet [16], TerraSegNet [17], ABCNet [18], class-feature-attention-based DeepLabv3+ [19], and self-attention/state-space-based lightweight segmentation networks [20], further demonstrate the effectiveness of global context modeling. Nevertheless, these methods mainly operate in the spatial domain. Full self-attention may introduce considerable computational cost for dense high-resolution imagery, whereas window-based or hierarchical attention may restrict long-distance information exchange. Moreover, spatial-domain attention alone does not explicitly exploit frequency-domain regularities in aerial scenes, such as repeated rooftops, road networks, vegetation textures, shadows, and large homogeneous regions.

Spatial-frequency representation learning provides a complementary perspective for remote sensing image segmentation. Frequency-domain information can describe global layouts, periodic textures, structural continuity, and boundary-related details beyond spatial convolution or spatial attention. Recent studies, including SFFNet [21], spatial–frequency dual-domain modeling [22], SF3Net [23], FSDENet [24], adaptive frequency enhancement [25], and frequency decoupling networks [26], have verified the potential of frequency-aware representation. However, many existing spatial-frequency methods still rely on direct concatenation, addition, fixed filtering, or single-stage frequency enhancement. These strategies may be insufficient for aligning spatial semantic features with frequency-domain structural cues, because the two domains describe image content from different perspectives. Therefore, an adaptive cross-domain fusion mechanism is needed to selectively integrate local spatial details and global frequency responses within an encoder–decoder segmentation framework.

Based on the above analysis, this study proposes a Multiscale Long-Distance Feature Aggregation Network, termed MLFANet, for semantic segmentation of high-resolution remote sensing imagery. The proposed model is designed to address three unresolved problems in existing segmentation paradigms. First, CNN-based multiscale methods improve local-to-global perception but often lack progressive cross-scale semantic propagation. Second, attention- and Transformer-based methods enhance long-distance spatial dependency modeling but usually ignore frequency-domain structural regularities. Third, existing

spatial-frequency methods exploit spectral cues but often lack adaptive alignment between spatial and frequency representations.

To address these problems, MLFANet introduces three key components. First, the Multiscale Global Dependency Extraction (MGDE) module performs cascaded multiscale contextual refinement by combining atrous convolution with polarized self-attention. Unlike conventional parallel multiscale aggregation, MGDE progressively propagates contextual responses across multiple receptive fields, thereby improving cross-scale semantic consistency. Second, an FFT-based frequency-domain branch with learnable global filtering is developed to capture global structural patterns and texture regularities that are difficult to explicitly represent in the spatial domain alone. Third, a bidirectional Spatial-Frequency Fusion (SFF) module is introduced to adaptively align local spatial details and global frequency responses through cross-attention, avoiding the semantic misalignment caused by simple feature concatenation.

The main contribution of this study is the proposal of MLFANet, a unified spatial-frequency feature aggregation framework for high-resolution remote sensing semantic segmentation. The proposed framework addresses scale variation, repeated textures, spatial heterogeneity, and long-distance semantic dependency through three key designs. First, the MGDE module performs cascaded multiscale contextual refinement by integrating atrous convolution with polarized self-attention, enabling progressive cross-scale semantic propagation. Second, the FFT-based frequency-domain branch introduces learnable global filtering to capture global structural patterns and texture regularities that are difficult to explicitly represent in the spatial domain alone. Third, the SFF module uses bidirectional cross-attention to adaptively align local spatial details and global frequency responses, reducing semantic misalignment between the two domains. Experimental results on the ISPRS Potsdam and Vaihingen datasets show that MLFANet achieves AF/MIoU/OA values of 86.03%/76.21%/88.70% and 83.17%/71.90%/86.33%, respectively, with 17.49 G FLOPs under a 256 × 256 input size. These results, together with ablation studies, verify the effectiveness and feasibility of the proposed components. Future work will further investigate lightweight model design, interpretable frequency-domain analysis, class-aware adaptive fusion, boundary-aware supervision, and cross-dataset generalization.

2 Related Works

2.1 CNN-Based Multiscale Feature Aggregation for Remote Sensing Segmentation

CNN-based encoder–decoder networks have been widely used in high-resolution remote sensing semantic segmentation because of their strong capacity for local feature extraction and hierarchical representation. In this paradigm, multiscale feature aggregation is usually achieved through feature pyramids, atrous convolution, pyramid pooling, multi-branch context encoding, or skip connections. Li et al. [9] proposed A2-FPN, which uses a dual-attention feature pyramid network to enhance spatial and channel-wise feature representation for fine-resolution remote sensing image segmentation. Zhou et al. [10] improved PSPNet by enhancing both global and local feature representations, thereby improving multiscale perception and boundary delineation. Chen et al. [11] proposed MCSNet, which integrates detailed local features and global contextual information through multi-branch encoding and contextual fusion. Li et al. [12] introduced MAResUNet, where multistage attention is embedded into a ResU-Net architecture to strengthen shallow and intermediate feature representations.

Although CNN-based multiscale aggregation methods have achieved strong performance, their feature fusion strategies are often based on parallel concatenation, addition, or fixed receptive-field expansion. Such designs are effective for capturing objects at different scales, but they may not sufficiently model semantic interactions across scales. This limitation is particularly important for high-resolution remote sensing imagery, where large objects such as buildings and impervious surfaces coexist with small objects

such as cars, and where objects from the same class may exhibit significant variations in shape, size, and texture. Therefore, beyond simply enlarging the receptive field, remote sensing segmentation requires a mechanism that can progressively refine and propagate semantic information across multiple spatial scales.

2.2 Attention and Transformer-Based Long-Range Dependency Modeling

Attention mechanisms and Transformer architectures have been introduced into remote sensing segmentation to overcome the locality limitation of conventional convolution. Li et al. [18] proposed ABCNet, an attentive bilateral contextual network that enhances contextual representation for efficient semantic segmentation of fine-resolution remote sensing imagery. He et al. [14] embedded Swin Transformer into a U-Net-like architecture and constructed STUNet to model hierarchical long-range interactions. Liu et al. [13] systematically rethought the use of Transformers for semantic segmentation of remote sensing images and showed that Transformer-based architectures require careful adaptation to remote sensing data characteristics. Wu et al. [15] proposed CMTFNet, which combines CNN and multiscale Transformer branches to integrate local spatial details and long-range semantic dependencies.

Recent studies have further explored attention-enhanced and hybrid CNN–Transformer architectures. For example, AFENet [16] introduces attention-focused feature enhancement for efficient remote sensing image segmentation, while TerraSegNet [17] uses bilateral axial attention and multiscale Transformer decoding for remote sensing segmentation under diverse environmental conditions. These methods demonstrate that attention-based models can improve global context modeling and semantic consistency. However, full self-attention remains computationally expensive for very high-resolution aerial images, while window-based attention may restrict information exchange across distant regions. More importantly, most attention and Transformer methods still operate mainly in the spatial domain. They usually model pixel-wise or token-wise similarity, but do not explicitly represent the spectral regularities, repeated textures, and global structural patterns that are common in aerial and satellite imagery.

2.3 Spatial-Frequency Representation Learning

Frequency-domain modeling has recently attracted increasing attention in remote sensing image interpretation. Compared with spatial-domain processing, frequency-domain analysis provides a complementary perspective for representing global structures, periodic textures, directional patterns, and high-frequency boundaries. In the context of remote sensing segmentation, Yang et al. [21] proposed SFFNet, a wavelet-based spatial and frequency domain fusion network, showing that frequency-domain cues can improve segmentation by complementing spatial features. Yao et al. [22] demonstrated that spatial–frequency dual-domain fusion can enhance remote sensing image representation by jointly exploiting spatial and spectral-domain information.

More recent works further confirm the usefulness of frequency-aware modeling. He et al. [23] proposed SF3Net, a frequency-domain enhanced segmentation network for high-resolution remote sensing imagery. Fu et al. [24] developed FSDENet, which enhances remote sensing semantic segmentation by integrating detail information from both frequency and spatial domains. Gao et al. [25] proposed an adaptive frequency enhancement network to improve semantic segmentation through frequency-domain feature enhancement. Li et al. [26] introduced a frequency decoupling network for remote sensing image segmentation, aiming to separate and exploit different frequency components for semantic representation.

Despite these advances, existing spatial-frequency methods still have several limitations. Some methods rely on simple concatenation, addition, or fixed filtering to combine spatial and frequency features, which may ignore the semantic discrepancy between the two domains. Other methods introduce frequency enhancement at a single stage but do not fully couple frequency-domain information with multiscale

encoder–decoder feature aggregation. In high-resolution remote sensing scenes, spatial features provide local geometry, object boundaries, and fine texture, while frequency features describe global regularities and structural patterns. These two types of representations should therefore be adaptively aligned rather than directly merged. This motivates the spatial-frequency cross-attention fusion strategy adopted in this study.

2.4 Summary of Research Gap

In summary, existing remote sensing semantic segmentation methods have advanced along several directions. However, these paradigms usually focus on only part of the problem. CNN-based methods are efficient in local and multiscale feature extraction, but they often lack explicit long-range dependency modeling. Transformer-based methods improve spatial global context modeling, but they remain mainly spatial-domain models and may introduce high computational cost. Spatial-frequency methods exploit frequency-domain cues, but many of them use simple fusion strategies and do not fully align frequency features with multiscale semantic representations. To make the research gap clearer, Table 1 compares the proposed MLFANet with representative existing paradigms from several key design dimensions.

Table 1: Experimental configuration information.

Method Paradigm	Multiscale Feature Aggregation	Explicit Long-Range Dependency Modeling	Frequency-Domain Representation	Adaptive Spatial-Frequency Fusion	Progressive Cross-Scale Semantic Propagation
CNN-based multiscale methods	✓	×	×	×	×
Attention based methods	✓	✓	×	×	△
Spatial-frequency methods	△	△	✓	△	×
Proposed MLFANet	✓	✓	✓	✓	✓

Note: ✓ indicates that the paradigm explicitly contains this design. △ indicates that the design is partially considered or indirectly achieved. × indicates that the design is generally absent or not the main focus.

3 Models and Methods

The proposed MLFANet follows an encoder-decoder architecture based on an improved ResNet50 backbone [27]. ResNet50 is adopted as the backbone because it provides a balanced hierarchical representation for high-resolution remote sensing segmentation. Compared with shallower ResNet variants, it offers stronger high-level semantic abstraction, while avoiding the excessive computational burden of deeper residual networks. The encoder generates hierarchical feature maps at four different spatial resolutions and removes the final fully connected layer to adapt to dense pixel-level prediction, as shown in Fig. 1.

To enhance cross-scale semantic understanding, the decoder incorporates a MGDE module and a frequency domain feature branch. The feature map $res4$ of the final ResNet bottleneck layer first passes through a 1×1 convolution layer, and then batch normalization and ReLU activation are carried out, as described in Eq. (1).

$$x_{Conv} = ReLU(BN\,(Conv(res4))) \tag{1}$$

This intermediate output is subsequently fed into the MGDE module and the frequency branch (FFT). The resulting features are subsequently fused using a cross-attention-based SFF, as described in Eq. (2):

$$x_{out} = SFF\,(MGDE((x_{Conv})\,, FFT\,(x_{Conv}))) \tag{2}$$

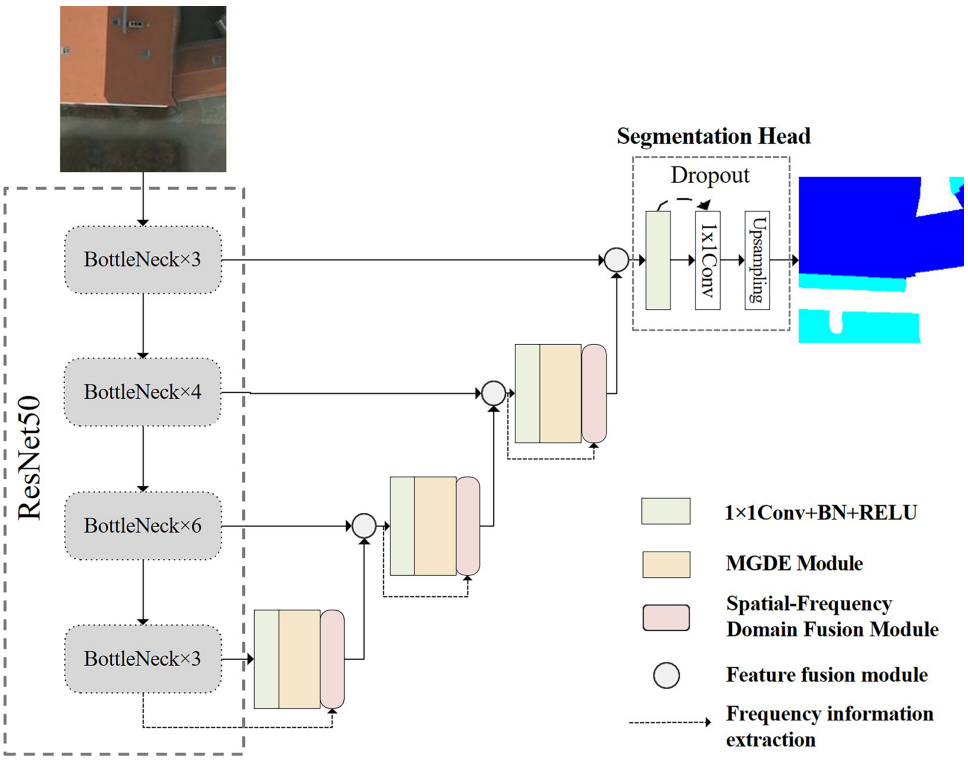

Figure 1: MLFANet framework.

To preserve low-level spatial details, the architecture employs U-Net-style skip connections. Features from each decoder stage are combined with corresponding encoder-stage features using a learnable weight coefficient α, generating the fused feature map F, as described in Eq. (3):

$$F = \alpha \cdot F_1 + (1 - \alpha) \cdot Upsample(F_2) \tag{3}$$

where F_1 represents the encoder feature and F_2 denotes the output from the fusion module. The segmentation head applies a 1×1 convolution convolution combined with bilinear interpolation to produce per-pixel predictions matching the input resolution.

3.1 Multiscale Global Dependency Extraction Module

In this network, the study proposes the MGDE module designed to enhance the model's ability to exploit spatial information across multiple scales in the encoder feature maps, while better capturing the long-range dependencies of objects at varying scales. Inspired by ASPP [28], the MGDE module performs dilated convolutions with different dilation rates and applies the Polarized Self-Attention (PSA) mechanism [29], as illustrated in Fig. 2.

The input features of MGDE are processed through four dilated convolution branches with different dilation rates, which can extract contextual information from multiple receptive fields. This multiscale representation is formulated in Eq. (4).

$$x_i = DilatedConv_i(x) \tag{4}$$

where x_i is the output of the i-th dilated convolution branch. Each branch output is refined through polarized self-attention mechanism (PSA), focusing on salient spatial and channel-wise features at each scale. MGDE

adopts sequential optimization instead of conventional parallel ASPP. The branch output is combined with the next branch input, and then processed by *PSA* to form the cascaded attention refinement pipeline, as described in Eq. (5).

$$z_{i+1} = PSA\left(z_i, x_{i+1}\right) \tag{5}$$

here, z_i denotes the refined feature from the i-th stage. This architecture enables each stage to iteratively refine and propagate feature significance, and helps fine-grained attention to the spatial structure. MGDE performs stacking and fuses the features of multi-branch outputs to enhance the expressive ability of the model. A 1×1 convolution adjusts the dimensionality of each output feature map to ensure compatibility for fusion with the corresponding encoder-level feature map, as defined in Eq. (6):

$$output = Conv\left(concat\left(z_1, z_2, z_3, z_4, global\left(x\right)\right)\right) \tag{6}$$

Compared to the parallel ASPP scheme, the proposed cascaded PSA-enhanced approach enables richer cross-scale interactions and boosts the network's capability to identify large objects by progressively integrating multi-resolution contextual information.

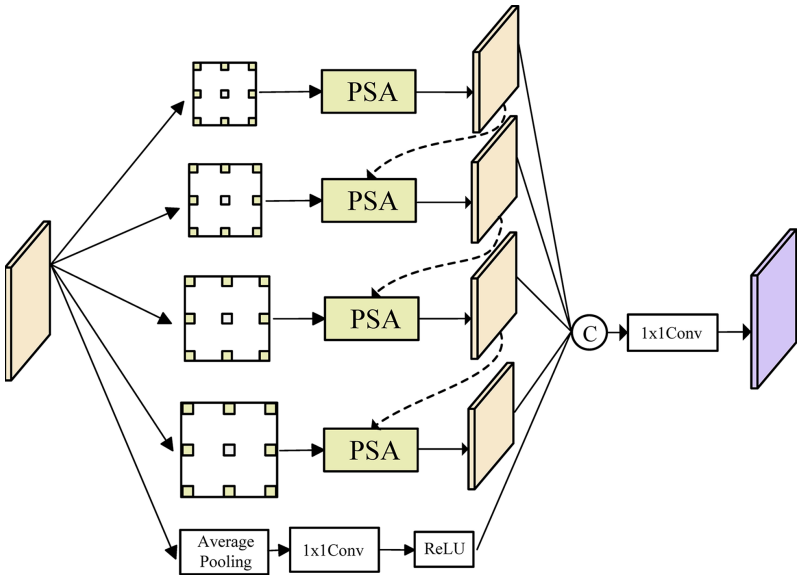

Figure 2: MGDE structure diagram.

3.2 Frequency Feature Extraction and Cross-Domain Fusion

To augment spatial representations with global semantic information, MLFANet includes a frequency feature extraction branch that captures global dependencies in the spectral domain, as illustrated in Fig. 3.

The input feature representation $x \in \mathbb{R}^{C \times H \times W}$ is first processed by convolution and normalization layers to enhance spatial encoding. A 2D Fast Fourier Transform (FFT) [30] is then applied to convert the representation into the frequency domain as illustrated in Eq. (7), where $\mathcal{F}[x]$ represents the 2D FFT. The output of the FFT is a complex-domain representation, which contains both amplitude and phase information of the input feature. The amplitude component reflects the strength distribution of different frequency responses, while the phase component preserves spatial structural information.

$$X = \mathcal{F}[x] \in \mathbb{C}^{H \times W \times D} \tag{7}$$

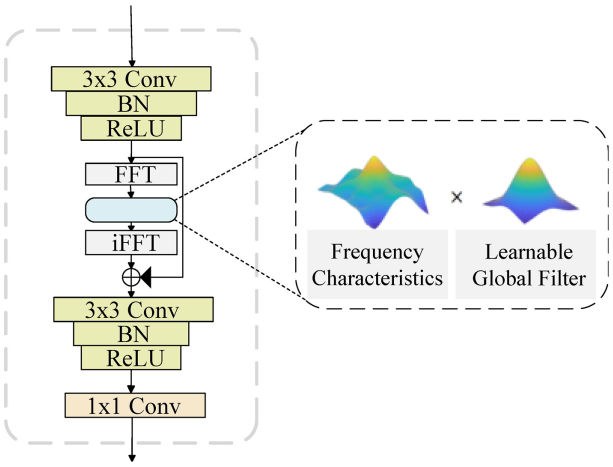

Figure 3: Frequency information extraction branch structure.

A set of learnable global filters $k \in \mathbb{R}^{C\times H\times W}$ modulates the frequency spectrum through element-wise multiplication. The modulated spectrum is the Hadamard product of the frequency domain information X and the global filter k, as shown in Eq. (8):

$$\tilde{X} = k \odot X \tag{8}$$

In this study, the learnable global filter is a real-valued parameter tensor with dimensions compatible with the frequency-domain feature. During the element-wise multiplication, the real-valued filter adaptively scales the corresponding complex frequency responses, thereby enhancing or suppressing different frequency components while preserving the complex-domain form required for the inverse FFT operation.

The filtered spectrum $\tilde{X}$ is converted back into the spatial domain via the inverse FFT, as expressed in Eq. (9):

$$x \leftarrow \mathcal{F}^{-1}[\tilde{X}] \tag{9}$$

After inverse transformation, the frequency-enhanced feature restores the same spatial resolution and channel dimension as the input feature of the frequency branch. Therefore, the output feature can be directly aligned with the spatial-domain feature and fed into the subsequent spatial-frequency fusion module.

To fuse spatial-frequency features, a cross-attention-based SFF is introduced, as shown in Fig. 4. The frequency-domain information captures global patterns and specific components within the feature map, and the spatial-domain information provides local details and structural information. By combining these complementary strengths, the model can perceive image features across multiple levels better. However, frequency-domain and spatial features represent distinct modalities that encode image attributes from different perspectives, and theoretically there are essential semantic differences. To address this challenge and achieve effective feature fusion, this paper proposes a spatial-frequency fusion module based on cross-attention, as shown in Fig. 4.

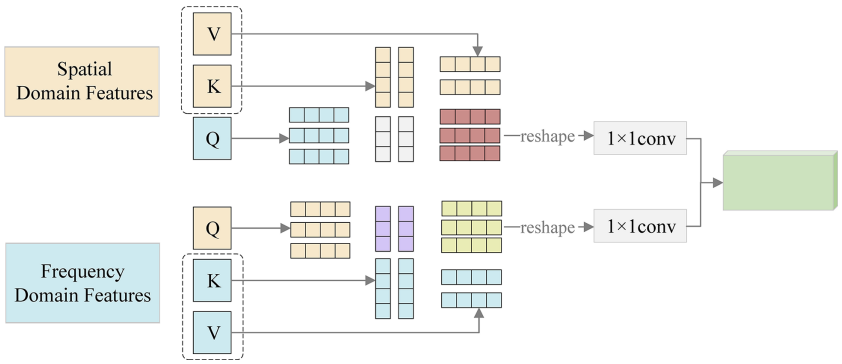

Figure 4: Spatial-frequency fusion module.

The spatial feature X_s and frequency feature X_f are independently projected into query (Q), key (K), and value (V) vectors. Cross-attention [31] is then computed in both directions. The spatial-to-frequency attention is formulated in Eq. (10):

$$A_{sf} = Softmax\left(\frac{Q_s K_f^T}{\sqrt{d_k}}\right), Z_{sf} = A_{sf} \cdot V_f \tag{10}$$

Similarly, the frequency-to-spatial attention calculation is shown in Eq. (11):

$$A_{fs} = Softmax\left(\frac{Q_f K_s^T}{\sqrt{d_k}}\right), Z_{fs} = A_{fs} \cdot V_s \tag{11}$$

The bidirectional results are summed to obtain the final fused representation, as shown in Eq. (12).

$$Z_{fusion} = Z_{sf} + Z_{fs} \tag{12}$$

The mechanism of cross-domain integration enables the network to utilize complementary spatial and spectral characteristics and enhance the ability to distinguish objects with subtle texture or structural variations in remote sensing imagery.

4 Results and Analysis

4.1 Dataset

This study evaluates the model using two high-resolution remote sensing datasets provided by the International Society for Photogrammetry and Remote Sensing (ISPRS), including Potsdam and Vaihingen. The Potsdam dataset comprises 38 ortho-rectified RGB-IR aerial images captured over Potsdam, Germany, as illustrated in Fig. 5. The spatial resolution of these images is 5 cm, and the size is 6000 × 6000 pixels. The data comprises red, green, blue, and near-infrared (NIR) bands, and the study utilizes data from three channels, NIR, red, and green, for training and testing. The dataset encompasses diverse urban environments, such as dense buildings, roads, vegetation, and vehicles. According to the standard partitioning method, there are 24 images in the training set and 14 images in the test set (https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx).

The Vaihingen dataset comprises 33 images acquired over Vaihingen, a village characterized by predominantly residential buildings and rural landscapes, as illustrated in Fig. 6.

(a) (b) (c)

Figure 5: ISPRS Potsdam dataset. (**a**) 38 regions of the Potsdam dataset; (**b**) Images of regions 2–10; (**c**) Labels of regions 2–10.

(a) (b) (c)

Figure 6: ISPRS Vaihingen dataset. (**a**) 33 regions of the Vaihingen dataset; (**b**) Images of region 1; (**c**) Labels of region 1.

Each image has 3 spectral bands, including NIR, red, and green. The average pixel size is 2494 × 2064, and the ground sampling distance (GSD) is 9 cm. Following the ISPRS data partition, 16 image tiles are designated for training purposes and 17 for testing. Given the negligible presence of clutter in this dataset, the clutter class is omitted from the evaluation metrics (https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx).

4.2 Evaluation Metrics

For assessing the semantic segmentation model's performance, five commonly used quantitative metrics are employed: Overall Accuracy (OA), F1-score, Mean Intersection over Union (MIoU), Intersection over Union (IoU), and Average F1-score (AF), with their formal definitions provided in Eqs. (13)–(17). These metrics are calculated based on the confusion matrix, which is made up of True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), and False Negative (*FN*).

(1) OA

$$\mathrm{OA} = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{13}$$

(2) F1-score (F1)

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

$$precision = \frac{TP_i}{TP_i + FP_i}, recall = \frac{TP_i}{TP_i + FN_i} \tag{14}$$

(3) IoU

$$\text{IoU} = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{15}$$

(4) MIoU

$$\text{MIoU} = \frac{1}{K} \sum_{i=1}^{K} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{16}$$

(5) AF

$$\text{AF} = \frac{1}{K} \sum_{i=1}^{K} F_i \tag{17}$$

where K denotes the number of evaluated semantic categories, and TP_i, FP_i, and FN_i represent the true positives, false positives, and false negatives of the i-th category, respectively. In this study, $K = 6$ for the Potsdam dataset and $K = 5$ for the Vaihingen dataset because the clutter class is excluded from the Vaihingen evaluation.

4.3 Implementation Details

The environment for experimentation configuration is detailed in Table 2.

Table 2: Experimental configuration information.

Experiments Involving Hardware/Software	Specific Configuration
CPU	i7-9800x
GPU	NVIDIA GeForce RTX 3090TI
Memory size	64 G
Deep learning frameworks	Pytorch 1.12
CUDA	11.6
Python	Python 3.10.9

On account of the high resolution of the original images, which exceeds the memory capacity of the laboratory's graphics card, each image is segmented into non-overlapping 256 × 256-pixel patches by means of a sliding window approach. Given that the number of training samples is sufficient, no data augmentation is applied in this study.

The hyperparameters employed in the experiments are tabulated in Table 3. The Adam optimizer is used for updating network parameters, and the learning rate undergoes dynamic adjustment through a cosine annealing strategy.

Table 3: Experimental parameter configuration.

Hyperparameters	Value
Batch size	32
Learning rate	0.0001
Epoch	Potsdam 60 Vaihingen 100
Optimizer	Adam
Weight decay	0.0001

This experiment employs the standard cross-entropy function as its loss function [32]. Let K denote the total number of categories to which image pixels can belong and let N represent the total quantity of pixels. Let $y_{i,c}$ indicate the true label of the i-th pixel for category c, and let $p_{i,c}$ denote the probability assigned by the model to the i-th pixel belonging to category c. The total cross-entropy loss is obtained through summation over all pixels and categories, with subsequent averaging, as defined in Eq. (18).

$$\mathrm{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} y_{i,c}\log\left(p_{i,c}\right) \tag{18}$$

4.4 Ablation Study

4.4.1 Effect of Backbone Networks

To analyze the influence of backbone selection on the proposed framework, four representative backbones were compared, including ResNet18, ResNet34, ResNet50, and Swin-Tiny. ResNet18 and ResNet34 were selected to evaluate the effect of shallower CNN backbones, ResNet50 was adopted as a deeper residual backbone with stronger semantic representation, and Swin-Tiny was included as a lightweight Transformer-based architecture to examine the suitability of self-attention-based feature extraction for the proposed network. All backbone variants were trained under the same experimental settings to ensure a fair comparison.

As shown in Tables 4 and 5, ResNet50 achieves the best overall performance on both ISPRS Potsdam and Vaihingen datasets. On the Potsdam dataset, ResNet50 obtains the highest AF, MIoU, and OA values, reaching 86.03%, 76.21%, and 88.70%, respectively. Compared with ResNet18 and ResNet34, ResNet50 provides stronger high-level semantic representation, which is beneficial for the proposed MGDE module to perform cross-scale dependency modeling and for the SFF module to integrate spatial and frequency-domain features. Although ResNet18 and ResNet34 have lower model complexity, their relatively shallow feature hierarchies may limit the representation of complex geospatial structures.

Table 4: Ablation experiment results of the backbone network on the ISPRS Potsdam dataset.

Backbone	F1						AF	MIoU	OA
	Imp.	Building	Low.	Tree	Car	Clutter			
ResNet18	90.22	94.95	84.15	83.37	88.15	70.48	85.22	74.99	88.12
ResNet34	90.89	95.13	84.41	83.05	88.25	70.14	85.31	75.18	88.39
Swin-Tiny	85.68	90.70	77.56	72.50	83.51	59.57	78.25	65.38	82.03
ResNet50	90.99	95.21	84.79	83.71	89.72	71.77	86.03	76.21	88.70

Table 5: Ablation experiment results of the backbone network on the ISPRS Vaihingen dataset.

Backbone	F1					AF	MIoU	OA
	Imp.	Building	Low.	Tree	Car			
ResNet18	85.87	91.48	73.01	86.17	72.27	81.76	69.86	85.19
ResNet34	85.22	91.29	71.62	85.89	67.05	80.21	67.94	84.59
Swin-Tiny	81.23	87.16	64.39	80.86	44.27	71.58	57.88	79.55
ResNet50	87.42	93.43	73.39	86.13	75.47	83.17	71.90	86.33

On the Vaihingen dataset, ResNet50 also achieves the best AF, MIoU, and OA values, reaching 83.17%, 71.90%, and 86.33%, respectively. Although ResNet18 slightly outperforms ResNet50 in the tree category, ResNet50 performs better in most other categories and shows stronger overall robustness. This indicates that a deeper residual backbone is more suitable for extracting discriminative semantic features in complex multi-class remote sensing scenes. Therefore, ResNet50 is selected as the backbone of MLFANet because it provides the best trade-off among semantic representation capability, training stability, and compatibility with the proposed multiscale and spatial-frequency feature aggregation modules.

4.4.2 Effect of MGDE Placement

This section investigates the contribution of the proposed MGDE module and determines its optimal position within the decoder. The experiments are carried out on the ISPRS Potsdam and Vaihingen datasets, and the effect of integrating the MGDE module at different decoder stages is assessed. In the ablation study, the insertion positions are marked as 1, 2 and 3. The larger the number, the closer it is to the final segmentation head. The results are presented in Table 6.

Table 6: Ablation experiment results of the MGDE module on the ISPRS Potsdam dataset.

MGDE Position	F1						AF	MIoU	OA
	Imp.	Building	Low.	Tree	Car	Clutter			
1	90.76	95.09	84.96	83.75	89.35	71.69	85.93	76.04	88.66
2	90.27	94.92	84.57	83.49	88.64	69.69	85.26	75.09	88.20
3	90.07	94.79	84.43	83.13	89.11	69.73	85.21	75.02	87.98
1+2	90.74	95.17	84.85	83.71	88.60	71.16	85.71	75.72	88.59
1+3	88.89	93.64	83.79	83.83	87.97	68.51	84.44	73.82	87.19
2+3	90.69	95.13	84.79	83.71	88.97	71.49	85.79	75.84	88.56
1+2+3	90.99	95.21	84.79	83.71	89.72	71.77	86.03	76.21	88.70

The results demonstrate that the placement of MGDE within the decoder has a notable influence on segmentation performance. Among all evaluated configurations, integrating MGDE at all three decoder stages (1+2+3) achieves the best overall performance in terms of AF, MIoU, and OA, whereas the 1+3 configuration yields the lowest overall results. In addition, the three-stage configuration obtains the best performance for most semantic categories. These results suggest that jointly integrating MGDE across all decoder stages can more effectively fuse multi-scale features and enhance the model's ability to capture contextual information.

As shown in Table 7, the placement of MGDE within the decoder has a clear influence on segmentation performance. Among all evaluated configurations, applying MGDE at all three decoder stages (1+2+3) achieves the best overall AF, MIoU, and OA results, whereas applying MGDE only at Position 2 yields the lowest overall performance. At the category level, the 1+2+3 configuration obtains the highest F1 scores for impervious surfaces, buildings, and cars, while the best results for low vegetation and trees are achieved by the 1+2 configuration and Position 1, respectively. These results indicate that integrating MGDE across all decoder stages generally provides the most effective overall configuration, although the optimal placement may vary for individual semantic categories.

Table 7: Ablation experiment results of the MGDE module on the ISPRS Vaihingen dataset.

MGDE Position	F1					AF	MIoU	OA
	Imp.	Building	Low.	Tree	Car			
1	85.75	91.54	73.39	86.47	73.71	82.17	70.39	85.36
2	85.97	91.82	72.59	86.19	71.99	81.71	69.85	85.28
3	86.88	92.90	72.76	86.19	75.06	82.76	71.31	85.93
1+2	86.48	92.26	73.48	86.44	72.19	82.17	70.49	85.81
1+3	86.78	92.60	73.26	86.03	73.11	82.36	70.76	85.78
2+3	86.83	92.66	73.40	86.06	72.09	82.21	70.59	85.90
1+2+3	87.42	93.43	73.39	86.13	75.47	83.17	71.90	86.33

However, for the low vegetation and tree classes, the highest F1 scores are not achieved with the all-stage configuration. Specifically, low vegetation attains optimal performance when MGDE is embedded only at position 1, whereas trees achieve peak results when MGDE is applied at positions 1 and 3. This behavior may stem from the imbalance in category proportions within high-spatial-resolution remote sensing imagery, where dominant categories like buildings and impervious surfaces occupy larger spatial extents. When MGDE is introduced across all stages, the model may prioritize these prevalent classes, thereby underemphasizing the finer, more localized features of minority classes like trees and low vegetation.

4.4.3 Effect of Core Components

To evaluate the contribution of each core component in MLFANet, ablation experiments are conducted on the ISPRS Potsdam and Vaihingen datasets. Three variants are compared with the full model: removing the MGDE module, removing the frequency information extraction module, and replacing the proposed spatial-frequency fusion module with simple feature concatenation. Table 8 presents the ablation results on the ISPRS Potsdam dataset. Table 9 reports the ablation results on the ISPRS Vaihingen dataset.

As shown in Table 8, the full MLFANet achieves the best AF and MIoU on the Potsdam dataset, reaching 86.03% and 76.21%, respectively. When the MGDE module is removed, AF and MIoU decrease to 85.75% and 75.79%, indicating that cascaded multiscale global dependency extraction is beneficial for improving contextual aggregation and cross-scale semantic consistency. Removing the frequency information extraction module decreases AF and MIoU to 85.88% and 75.96%, respectively. This shows that the FFT-based branch provides complementary spectral cues for representing global structures and repeated textures beyond spatial-domain aggregation. When the spatial-frequency fusion module is replaced with simple concatenation, MIoU also decreases to 75.97%, suggesting that direct feature merging is less effective than adaptive cross-domain alignment. These results verify that the proposed modules jointly improve segmentation performance on complex urban aerial scenes.

As shown in Table 9, the full model also achieves the best overall performance on the Vaihingen dataset, with AF, MIoU, and OA values of 83.17%, 71.90%, and 86.33%, respectively. Compared with the full model, removing MGDE decreases MIoU by 1.05%, which confirms the importance of multiscale semantic refinement for scenes containing objects with different sizes and irregular spatial distributions. Removing the frequency information extraction module causes a more evident performance decline, with AF, MIoU, and OA decreasing by 1.68%, 2.36%, and 1.09%, respectively. This result empirically supports the necessity of frequency-domain modeling, especially for Vaihingen scenes that contain complex residential layouts, vegetation textures, shadows, and fine object boundaries. These characteristics require not only spatial-domain context but also frequency-domain cues related to global structural regularities and high-frequency texture details.

Table 8: Ablation results of core components on the ISPRS Potsdam dataset.

Removed Module	AF	MIoU	OA
MGDE	85.75	75.79	88.55
Frequency information extraction module	85.88	75.96	88.72
Spatial-frequency domain fusion module	85.88	75.97	88.63

Table 9: Results of ablation experiments on the ISPRS Vaihingen dataset.

Removed Module	AF	MIoU	OA
MGDE	82.43	70.85	85.81
Frequency information extraction module	81.49	69.54	85.24
Spatial-frequency domain fusion module	81.70	69.80	85.22
Full model proposed	83.17	71.90	86.33

Replacing the proposed spatial-frequency fusion module with simple concatenation also leads to clear degradation. AF, MIoU, and OA decrease to 81.70%, 69.80%, and 85.22%, respectively. This indicates that spatial and frequency features should not be directly merged without alignment, because they describe image content from different representational perspectives. The proposed bidirectional cross-attention fusion adaptively integrates local spatial details and global frequency responses, thereby improving the discriminative representation of high-resolution remote sensing scenes. Overall, the ablation results on both datasets demonstrate that MGDE, frequency information extraction, and spatial-frequency fusion all contribute to the final performance of MLFANet.

4.5 Comparison Experiments

To evaluate the performance of the proposed MLFANet, a series of comparative experiments is conducted with several representative and advanced semantic segmentation models, including ABCNet [18], PSPNet [10], MCSNet [11], MAResUNet [12], CMTFNet [15], and the proposed MLFANet. The compared methods are selected to cover the main methodological paradigms related to this study, including CNN-based segmentation, multiscale context aggregation, attention-enhanced feature modeling, and CNN–Transformer hybrid segmentation. This selection is consistent with the objective of this work, which is to evaluate the effectiveness of the proposed dual-domain feature aggregation paradigm rather than to construct an exhaustive benchmark leaderboard. All compared models are trained under the same hyperparameter

settings to ensure a fair comparison within this evaluation protocol. Quantitative results on the ISPRS Potsdam dataset are summarized in Table 10.

Table 10: Comparative experimental results of different network models on the ISPRS Potsdam dataset.

MGDE Position	F1						AF	MIoU	OA
	Imp.	Building	Low.	Tree	Car	Clutter			
ABCNet	90.11	94.77	83.92	83.03	86.74	70.05	84.77	74.30	87.88
PSPNet	90.63	95.17	84.43	83.33	85.78	70.65	84.99	74.64	88.31
MCSNet	87.46	92.69	79.72	76.01	84.00	61.39	80.21	68.65	84.12
MAResUNet	90.19	94.89	84.03	83.19	87.22	69.17	84.78	74.38	87.96
CMTFNet	90.79	95.19	84.79	83.59	89.99	71.18	85.92	76.07	88.59
MLFANet	90.99	95.21	84.79	83.71	89.72	71.77	86.03	76.21	88.70

The experimental results on the ISPRS Potsdam dataset show that MLFANet achieves competitive overall performance among the compared methods. In terms of the three overall evaluation metrics, MLFANet obtains the best AF, MIoU, and OA values, reaching 86.03%, 76.21%, and 88.70%, respectively. Compared with the strongest competing model, CMTFNet, the improvements in AF, MIoU, and OA are 0.11%, 0.14%, and 0.11%, respectively. These gains are relatively limited, indicating that MLFANet provides moderate rather than substantial improvement on the Potsdam dataset. For class-wise performance, MLFANet achieves the best or tied-best F1-score in most categories, including impervious surfaces, buildings, low vegetation, trees, and clutter. The largest class-wise gain over CMTFNet appears in the clutter category, with an improvement of 0.59%. However, the F1-score for cars is 0.27% lower than that of CMTFNet, indicating that small-object segmentation remains challenging. For a further evaluation of visual performance, qualitative comparisons on the Potsdam dataset are presented in Fig. 7.

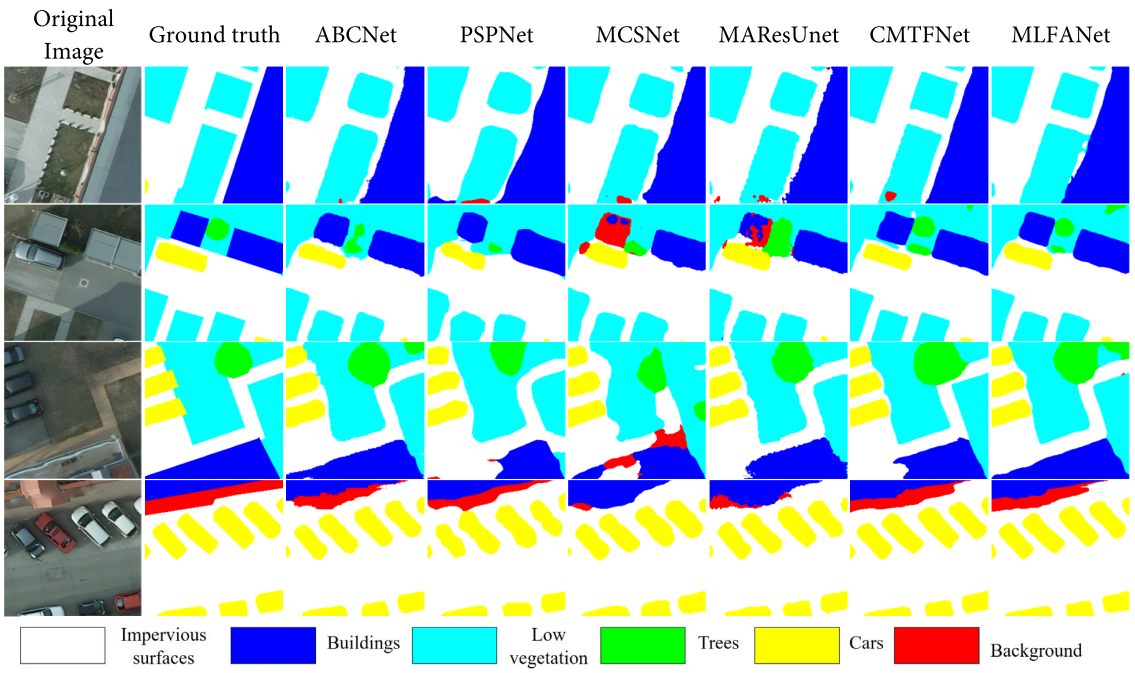

Figure 7: Qualitative comparison between different methods on the ISPRS Potsdam dataset.

In the first row of the visual results, most baseline models exhibit varying degrees of confusion between low vegetation and the background, whereas MLFANet accurately differentiates these two classes. In the second and third rows, MLFANet produces more complete and smoother building boundaries, reducing misclassification into the background and enabling more accurate contour delineation. Additionally, in the third row, MLFANet segments roads with greater continuity and coherence. In the fourth row, both MLFANet and CMTFNet achieve similar performance in distinguishing between background and buildings, demonstrating comparable robustness in this particular scenario.

Table 11 presents the comparison results on the Vaihingen dataset. MLFANet attains the best performance across overall metrics and four out of five semantic classes, with average F1-score, MIoU, and OA reaching 83.17%, 71.90%, and 86.33%, respectively. The only exception is the tree class, where MAResUNet attains a marginally higher F1-score, which is 0.1% above that of MLFANet. The superiority of MLFANet is especially pronounced for large-scale categories including impervious surfaces and buildings, demonstrating consistently balanced segmentation performance across all classes. These findings additionally confirm the model's capability in managing multiscale object segmentation tasks in high-resolution aerial imagery.

Table 11: Comparative experimental results of different network models on the ISPRS Vaihingen dataset.

Model	F1					AF	MIoU	OA
	Impervious Surfaces	Buildings	Low Vegetation	Trees	Cars			
ABCNet	85.15	91.19	70.78	85.32	71.94	80.88	68.66	84.42
PSPNet	85.91	92.09	72.04	85.44	68.19	80.73	68.65	85.06
MCSNet	81.35	88.06	64.68	82.72	47.15	72.79	59.28	80.44
MAResUNet	86.31	92.13	73.05	86.23	75.29	82.60	71.47	85.39
CMTFNet	86.19	92.17	73.22	86.01	75.29	82.58	71.33	85.44
MLFANet	87.42	93.43	73.39	86.13	75.47	83.17	71.90	86.33

To visualize fine-grained segmentation results on the Vaihingen dataset, the study presents detailed outputs in Fig. 8.

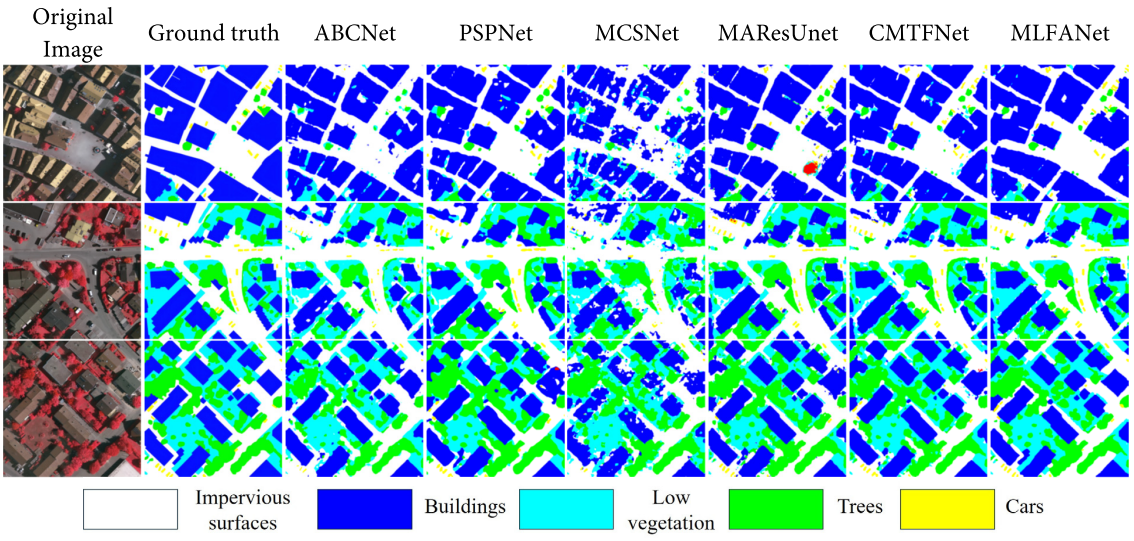

Figure 8: Qualitative comparison of different methods on the ISPRS Vaihingen dataset.

Fig. 9 specifically magnifies the bottom-right portion of the second row in Fig. 8, providing a comparative analysis of the segmentation performance between CMTFNet and MLFANet on a representative sub-region. This area includes a large road segment with sparsely distributed vehicles and encompasses all annotated classes in the dataset, thereby providing a thorough evaluation of model behavior.

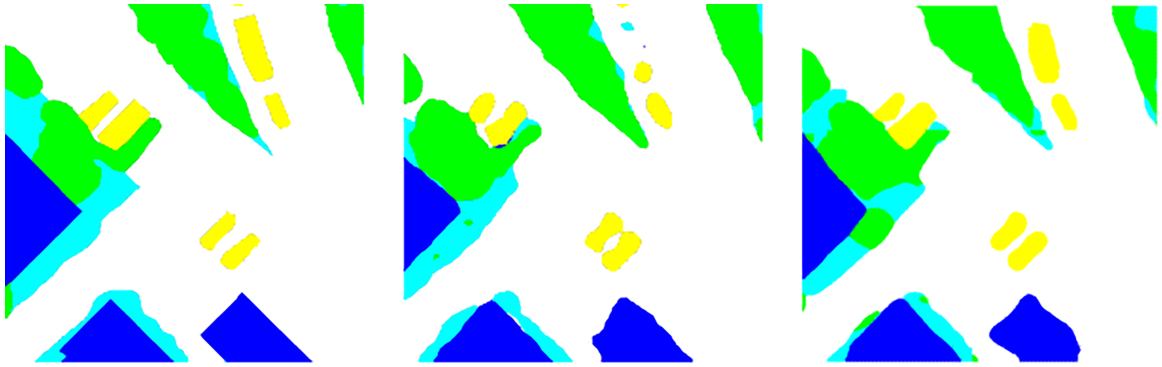

Figure 9: Locally enlarged view of the visualization results of the Vaihingen dataset; (**a**) Ground truth label; (**b**) Segmentation result of CMTFNet; (**c**) Segmentation result of MLFANet.

Fig. 9a displays the ground truth, Fig. 9b presents the segmentation output of CMTFNet, and Fig. 9c shows the result produced by MLFANet. While CMTFNet correctly delineates most buildings, it fails to detect the small car in the upper-right corner and exhibits misclassifications between trees and cars. In contrast, MLFANet accurately segments both cars and buildings, with only minor errors occurring in overlapping regions of trees and low vegetation. These results demonstrate MLFANet's enhanced capability in capturing multiscale features and its superior performance in segmenting both large- and small-scale objects within complex remote sensing scenes.

From the qualitative results, the differences among the compared models are distributed across the entire segmentation map rather than being limited to a single localized region. The main visual differences are reflected in building boundary continuity, regional consistency of impervious surfaces, fragmentation of vegetation areas, recognition of small objects such as cars, and confusion between low vegetation and trees. MLFANet produces more complete object regions and more consistent semantic predictions in complex urban scenes. It better preserves building structures and reduces missed detections of small vehicles compared with several baseline methods. However, some errors still remain in areas where low vegetation and trees have similar spectral and spatial characteristics, especially near shadows or ambiguous boundaries.

4.6 Computational Complexity Analysis

To further evaluate the practical applicability of the proposed MLFANet, the computational complexity of different segmentation models is compared in terms of the number of trainable parameters and floating-point operations (FLOPs). All models are evaluated under the same input size of 256 × 256 pixels, which is consistent with the patch size used in the experiments. The results are shown in Table 12.

As shown in Table 12, MLFANet contains 43.90 M parameters and requires 17.49 G FLOPs. Although the number of parameters is larger than that of ABCNet, MAResUNet, and CMTFNet, its computational cost remains relatively low. Specifically, the FLOPs of MLFANet are close to those of MCSNet and substantially

lower than those of PSPNet, MAResUNet, and CMTFNet. This indicates that the proposed MGDE, FFT-based frequency branch, and spatial-frequency fusion module increase the representational capacity of the model without causing excessive computational overhead.

Table 12: Computational complexity comparison of representative segmentation models.

Method	Params (M)	FLOPs (G)
ABCNet	13.39	15.58
PSPNet	53.32	28.00
MCSNet	42.41	17.43
MAResUNet	26.28	35.09
CMTFNet	30.07	32.85
MLFANet	43.90	17.49

Compared with CMTFNet, MLFANet achieves better segmentation performance while reducing FLOPs from 32.85 to 17.49 G. This suggests that the proposed spatial-frequency feature aggregation strategy provides a favorable trade-off between accuracy and computational efficiency.

5 Discussion

This study proposes MLFANet to address several intrinsic challenges in high-resolution remote sensing semantic segmentation, including severe scale variation, long-distance semantic dependency, repeated textures, and spatial heterogeneity. Different from methods that mainly rely on stronger backbones or isolated attention modules, MLFANet follows a unified feature aggregation paradigm that combines progressive cross-scale semantic propagation, frequency-domain global representation, and adaptive spatial-frequency fusion.

The MGDE module is designed to improve multiscale dependency modeling. Existing CNN-based methods often aggregate multiscale features through parallel atrous convolution, pyramid pooling, or direct feature concatenation. Although these strategies enlarge the receptive field, they do not explicitly propagate semantic information across different scales. In contrast, MGDE performs cascaded multiscale refinement with polarized self-attention, enabling features from different receptive fields to interact progressively. The ablation results show that removing MGDE reduces the performance on both Potsdam and Vaihingen datasets, confirming its contribution to cross-scale semantic modeling.

The FFT-based frequency branch provides complementary global structural information. High-resolution aerial scenes often contain repeated rooftops, road networks, vegetation textures, shadows, and large homogeneous regions. These patterns are not always sufficiently captured by spatial convolution or spatial attention alone. Frequency-domain features can represent low-frequency regional consistency and high-frequency boundary or texture details. The ablation results indicate that removing the frequency branch leads to performance degradation, especially on the Vaihingen dataset. This supports the necessity of introducing frequency-domain modeling into remote sensing segmentation.

The SFF module further improves feature integration by adaptively aligning spatial and frequency representations. Since spatial features emphasize local geometry and semantic details, while frequency features describe global texture and structural distributions, simple concatenation may cause semantic misalignment. The proposed bidirectional cross-attention fusion enables the two domains to interact more selectively. Compared with direct fusion, SFF improves AF, MIoU, and OA on both datasets, showing that

the gain of MLFANet comes not only from adding a frequency branch, but also from effective dual-domain feature alignment.

The comparison experiments show that MLFANet achieves competitive performance against representative segmentation networks. On the Potsdam dataset, MLFANet obtains the best AF, MIoU, and OA, although the improvement over CMTFNet is moderate rather than large. On the Vaihingen dataset, MLFANet shows clearer advantages in overall metrics and most semantic categories. Qualitative results also show that MLFANet produces more complete building regions, smoother boundaries, and better small-object segmentation in several local regions, while errors still occur around vegetation overlap and ambiguous boundaries.

In terms of computational complexity, MLFANet contains 43.90 M parameters and 17.49 G FLOPs. Although its parameter size is larger than several lightweight baselines, its FLOPs remain lower than PSPNet, MAResUNet, and CMTFNet. This suggests that MLFANet improves feature representation while maintaining a moderate computational cost. However, it should not be regarded as a lightweight model, and the relatively large parameter size may limit deployment in real-time or resource-constrained scenarios.

Despite these promising results, several limitations remain. First, MLFANet introduces additional parameters and computational cost due to the use of multiscale attention refinement, FFT-based filtering, and bidirectional spatial-frequency fusion. Therefore, future work will focus on lightweight network design and efficient frequency filtering to improve deployment feasibility. Second, although the frequency-domain branch is quantitatively verified through ablation experiments, the learned frequency responses are not explicitly visualized in this study. Since the FFT-based branch is optimized in an end-to-end manner together with learnable global filters and the SFF module, its intermediate responses may not directly correspond to manually interpretable land-cover categories. Future work will investigate more interpretable frequency-response visualization, class-aware spectral analysis, and attribution-based methods to better explain how frequency cues contribute to different semantic classes. Third, the current model still shows confusion between spectrally and spatially similar classes, such as low vegetation and trees, especially near ambiguous boundaries or shadowed regions. Incorporating class-aware adaptive fusion, boundary-aware supervision, and additional multispectral or multi-temporal information may further improve the discrimination of such challenging categories. Finally, the experiments are conducted on two standard ISPRS benchmark datasets, and future studies will evaluate the generalization ability of MLFANet on larger-scale and more diverse remote sensing datasets.

6 Conclusions

This study proposed MLFANet, a multiscale long-distance feature aggregation network for semantic segmentation of high-resolution remote sensing imagery. The motivation of MLFANet is to address the limitations of existing segmentation paradigms, where CNN-based methods mainly focus on local and multiscale spatial aggregation, attention- or Transformer-based methods emphasize spatial long-range dependency modeling, and spatial-frequency methods often lack adaptive semantic alignment between the two domains. To bridge these gaps, MLFANet integrates cascaded cross-scale semantic propagation, FFT-based frequency-domain representation, and bidirectional spatial-frequency fusion within a unified encoder–decoder framework. Specifically, the MGDE module progressively refines multiscale contextual features and strengthens cross-scale dependency modeling. The frequency-domain branch introduces learnable FFT-based filtering to capture global structural regularities and texture patterns that are difficult to explicitly represent in the spatial domain alone. The SFF module further aligns spatial and frequency features through bidirectional cross-attention, enabling adaptive integration of local details and global frequency responses. Experiments on the ISPRS Potsdam and Vaihingen datasets show that MLFANet achieves competitive

performance compared with representative segmentation models. Ablation studies confirm the effectiveness of the proposed MGDE, frequency-domain modeling, and spatial-frequency fusion components.

Nevertheless, several limitations remain. MLFANet still has a relatively large number of parameters due to the ResNet50 backbone and additional refinement modules. In addition, the improvement is more evident for large and structurally coherent categories, while confusion between low vegetation and trees may still occur because of similar spectral responses and irregular textures. The current evaluation is also limited to two benchmark datasets, and further validation on more diverse sensors, seasons, and geographic regions is needed. Future work will focus on lightweight network design, class-aware adaptive fusion, efficient frequency filtering, and boundary-aware supervision to improve both segmentation accuracy and deployment efficiency in complex geospatial scenarios.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Guangyu Xu and Wenfeng Zheng; methodology, Guangyu Xu and Yuxi Ban; software, Guangyu Xu and Legend Zhang; validation, Guangyu Xu, Yuxi Ban and Junmin Lyu; formal analysis, Guangyu Xu and Junmin Lyu; investigation, Guangyu Xu and Yuxi Ban; resources, Feng Bao and Wenfeng Zheng; data curation, Guangyu Xu and Legend Zhang; writing—original draft preparation, Guangyu Xu; writing—review and editing, Yuxi Ban, Legend Zhang, Junmin Lyu, Feng Bao and Wenfeng Zheng; visualization, Guangyu Xu and Legend Zhang; supervision, Feng Bao and Wenfeng Zheng; project administration, Wenfeng Zheng. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available from the International Society for Photogrammetry and Remote Sensing (ISPRS) 2D Semantic Labeling Benchmark. The Potsdam dataset is available at https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx, and the Vaihingen dataset is available at https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx.

Ethics Approval: Not applicable.

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

References

1. Kotaridis I, Lazaridou M. Remote sensing image segmentation advances: a meta-analysis. ISPRS J Photogramm Remote Sens. 2021;173(3):309–22. doi:10.1016/j.isprsjprs.2021.01.020.
2. Yuan X, Shi J, Gu L. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst Appl. 2021;169(6214):114417. doi:10.1016/j.eswa.2020.114417.
3. Li J, Cai Y, Li Q, Kou M, Zhang T. A review of remote sensing image segmentation by deep learning methods. Int J Digit Earth. 2024;17(1):2328827. doi:10.1080/17538947.2024.2328827.
4. Hoeser T, Kuenzer C. Object detection and image segmentation with deep learning on earth observation data: a review—part I: evolution and recent trends. Remote Sens. 2020;12(10):1667. doi:10.3390/rs12101667.
5. Huang L, Jiang B, Lv S, Liu Y, Fu Y. Deep-learning-based semantic segmentation of remote sensing images: a survey. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:8370–96. doi:10.1109/JSTARS.2023.3335891.
6. Lv J, Shen Q, Lv M, Li Y, Shi L, Zhang P. Deep learning-based semantic segmentation of remote sensing images: a review. Front Ecol Evol. 2023;11:1201125. doi:10.3389/fevo.2023.1201125.
7. Xie Y, Zhan N, Zhu J, Xu B, Chen H, Mao W, et al. Landslide extraction from aerial imagery considering context association characteristics. Int J Appl Earth Obs Geoinf. 2024;131(11):103950. doi:10.1016/j.jag.2024.103950.

8. Hui J, Mi W, Wang J, Cao Y, Zhou Z, Zheng N. AFSIFormer: adaptive frequency-spatial interaction attention mechanism for aerial image semantic segmentation. IEEE Trans Geosci Remote Sens. 2025;63:1–19. doi:10.1109/tgrs.2025.3599214.
9. Li R, Wang L, Zhang C, Duan C, Zheng S. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. Int J Remote Sens. 2022;43(3):1131–55. doi:10.1080/01431161.2022.2030071.
10. Zhou W, Wang C, Zhang X. Remote sensing image semantic segmentation algorithm based on improved PSPNet. In: Proceedings of the 2025 4th International Conference on Image Processing and Media Computing (ICIPMC); 2025 Jun 27–29; Xi'an, China. p. 15–21.
11. Chen Y, Wang Y, Xiong S, Lu X, Zhu XX, Mou L. Integrating detailed features and global contexts for semantic segmentation in ultrahigh-resolution remote sensing images. IEEE Trans Geosci Remote Sens. 2024;62(11):4703914. doi:10.1109/TGRS.2024.3394449.
12. Li R, Zheng S, Duan C, Su J, Zhang C. Multistage attention ResU-net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci Remote Sens Lett. 2022;19:8009205. doi:10.1109/LGRS.2021.3063381.
13. Liu Y, Zhang Y, Wang Y, Mei S. Rethinking transformers for semantic segmentation of remote sensing images. IEEE Trans Geosci Remote Sens. 2023;61:5617515. doi:10.1109/TGRS.2023.3302024.
14. He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens. 2022;60:4408715. doi:10.1109/TGRS.2022.3144165.
15. Wu H, Huang P, Zhang M, Tang W, Yu X. CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation. IEEE Trans Geosci Remote Sens. 2023;61:2004612. doi:10.1109/TGRS.2023.3314641.
16. Li J, Cheng S. AFENet: an attention-focused feature enhancement network for the efficient semantic segmentation of remote sensing images. Remote Sens. 2024;16(23):4392. doi:10.3390/rs16234392.
17. Setyawan Wijaya B, Munir R, Utama NP. TerraSegNet: bilateral axial attention network for remote sensing image segmentation in diverse environmental monitoring applications. IEEE Access. 2025;13:208868–901. doi:10.1109/ACCESS.2025.3633712.
18. Li R, Zheng S, Zhang C, Duan C, Wang L, Atkinson PM. ABCNet: attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. ISPRS J Photogramm Remote Sens. 2021;181(12):84–98. doi:10.1016/j.isprsjprs.2021.09.005.
19. Wang Z, Wang J, Yang K, Wang L, Su F, Chen X. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. Comput Geosci. 2022;158:104969. doi:10.1016/j.cageo.2021.104969.
20. Li L, Yi J, Fan H, Lin H. A lightweight semantic segmentation network based on self-attention mechanism and state space model for efficient urban scene segmentation. IEEE Trans Geosci Remote Sens. 2025;63:4703215. doi:10.1109/TGRS.2025.3562185.
21. Yang Y, Yuan G, Li J. SFFNet: a wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. IEEE Trans Geosci Remote Sens. 2024;62:3000617. doi:10.1109/TGRS.2024.3427370.
22. Yao Z, Fan G, Fan J, Gan M, Philip Chen CL. Spatial-frequency dual-domain feature fusion network for low-light remote sensing image enhancement. IEEE Trans Geosci Remote Sens. 2024;62:1–16. doi:10.1109/tgrs.2024.3434416.
23. He Y, Lu Z, Huan H. SF3Net: frequency-domain enhanced segmentation network for high-resolution remote sensing imagery. Remote Sens. 2025;17(22):3734. doi:10.3390/rs17223734.
24. Fu J, Yu Y, Wang L. FSDENet: a frequency and spatial domains-based detail enhancement network for remote sensing semantic segmentation. IEEE J Sel Top Appl Earth Obs Remote Sens. 2025;18:19378–92. doi:10.1109/JSTARS.2025.3583558.
25. Gao F, Fu M, Cao J, Dong J, Du Q. Adaptive frequency enhancement network for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens. 2025;63:5619415. doi:10.1109/TGRS.2025.3558472.
26. Li X, Xu F, Yu A, Lyu X, Gao H, Zhou J. A frequency decoupling network for semantic segmentation of remote sensing images. IEEE Trans Geosci Remote Sens. 2025;63:5607921. doi:10.1109/TGRS.2025.3531879.
27. 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); 2016 Jun 27–30; Las Vegas, NV, USA.

28. Srivastava N, Rai A, Prasad Kushwaha SK, Jain K. Advancing multi-class semantic segmentation of high-resolution satellite imagery through enhanced ASPP and attention mechanisms. In: Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium; 2024 Jul 7–12; Athens, Greece. p. 8423–7.
29. Liu H, Liu F, Fan X, Huang D. Polarized self-attention: towards high-quality pixel-wise mapping. Neurocomputing. 2022;506(4):158–67. doi:10.1016/j.neucom.2022.07.054.
30. Cooley JW, Tukey JW. An algorithm for the machine calculation of complex Fourier series. Math Comput. 1965;19(90):297–301. doi:10.1090/s0025-5718-1965-0178586-1.
31. Ma J, Jiang W, Tang X, Zhang X, Liu F, Jiao L. Multiscale sparse cross-attention network for remote sensing scene classification. IEEE Trans Geosci Remote Sens. 2025;63:5605416. doi:10.1109/TGRS.2025.3525582.
32. Mao A, Mohri M, Zhong Y. Cross-entropy loss functions: theoretical analysis and applications. In: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA.