remote sensing

MDPI

Article

Context-Aware Feature Enhancement Network for Remote Sensing Image Semantic Segmentation

Shufen Ruan [1,2], Quan Wan [1], Ruijuan Chen [1,2], Mengyang Hu [1], Xiuya Guo [1,2] and Kunfang Song [3,*]

[1] School of Mathematics and Statistics, Wuhan Textile University, Wuhan 430200, China; sfruan@wtu.edu.cn (S.R.); 2315173007@wtu.edu.cn (Q.W.); rjchen@wtu.edu.cn (R.C.); 2415173014@wtu.edu.cn (M.H.); xyguo@wtu.edu.cn (X.G.)
[2] Research Center for Applied Mathematics and Interdisciplinary Sciences, Wuhan Textile University, Wuhan 430200, China
[3] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China
* Correspondence: skf@wtu.edu.cn

Highlights

What are the main findings?

- We propose a context-aware semantic segmentation framework, named CFENet, which integrates the WPPM, ACE, and MFR modules designed in the manuscript and achieves outstanding segmentation performance on both the Vaihingen and Potsdam datasets.
- The WPPM employs a wavelet-based scale reorganization strategy to enhance contextual perception. The ACE module enables adaptive interaction between spatial and channel information, which significantly improves multi-scale feature representation. Meanwhile, the MFR utilizes a concise convolutional design for efficient feature decoding.

What are the implications of the main findings?

- CFENet provides an effective and reliable approach for high-accuracy semantic segmentation of remote sensing images.
- The modular design of the CFENet system enhances extensibility and applicability, which can be conveniently integrated into various encoder-decoder architectures, providing an efficient scheme for downstream remote sensing applications, such as land-cover mapping and environmental monitoring.

Abstract

Semantic segmentation of remote sensing images plays a crucial role in accurate land-cover classification and environmental monitoring. However, existing semantic segmentation networks still struggle with multiscale feature extraction and context modeling. To address these challenges, this paper proposes a novel semantic segmentation network, termed Context-aware Feature Enhancement Network (CFENet). Specifically, we design a Wavelet-Based Pyramid Pooling Module (WPPM) based on Haar wavelet downsampling (HWD) to enhance the model's ability to extract multiscale features. Meanwhile, an Adaptive Context Enhancement (ACE) module is introduced to adaptively focus on semantically significant regions, enabling joint enhancement along both spatial and channel dimensions. In addition, we develop a Multiscale Feature Reconstruction (MFR) module that performs multiscale decoding on the output of ACE in the decoding stage to further improve segmentation accuracy. The effectiveness of CFENet is validated on two benchmark datasets: ISPRS Vaihingen and ISPRS Potsdam. Experimental results show that, compared to baseline models, CFENet improves the mF1 by 3.16% and 2.38%, the OA by 1.54% and 2.80%, and the mIoU by 4.44% and 3.91%, respectively. Moreover, CFENet achieves reliable and satisfactory performance when evaluated against several representative mainstream methods.

Check for updates

Academic Editor: Tao Lei

Received: 24 December 2025
Revised: 4 February 2026
Accepted: 5 February 2026
Published: 8 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: semantic segmentation; remote sensing image; multiscale feature extraction; context modeling

1. Introduction

Semantic segmentation assigns a class label to each pixel, enabling fine-grained understanding of visual content. It plays a crucial role in various fields such as industrial inspection [1], autonomous driving [2,3], remote sensing [4], and medical imaging [5]. Among various application scenarios, semantic segmentation of remote sensing images has gradually attracted extensive attention from researchers due to its unique data characteristics and practical demands. Compared with semantic segmentation of natural images, remote sensing imagery often features large spatial dimensions, complex textures, and high intra-class variability, which pose significant challenges for accurate pixel-level classification [6].

Early methods for remote sensing image semantic segmentation mainly employed traditional image processing and shallow machine learning models, such as random forests [7], support vector machines [8], and conditional random fields (CRFs) [9]. These methods suffer from strong dependency on hand-crafted feature representations and limited generalization ability.

With the advent of deep learning, FCN [10] and U-Net [11] pioneered end-to-end CNN-based semantic segmentation and laid the foundation for remote sensing image segmentation. Mnih and Hinton first demonstrated the effectiveness of CNNs for labeling high-resolution aerial imagery, pioneering pixel-level interpretation in remote sensing scenes [12]. Sherrah adopted fully convolutional architectures for urban scene labeling in aerial images, enabling end-to-end semantic segmentation while preserving spatial resolution [13]. To further improve spatial detail recovery and computational efficiency, Maggiori et al. proposed a convolutional neural network framework tailored for high-resolution remote sensing imagery, which effectively addressed complex urban structures [14]. Later, multi-scale feature extraction strategies were introduced to handle the significant scale variations of objects in remote sensing images, enhancing the segmentation performance of buildings, vegetation, and roads [15].

At present, most networks employ Transformer modules and attention mechanisms to model global contextual relationships. DC-Swin [16] introduces Transformer as the backbone to extract contextual information, achieving good performance in remote sensing scenarios. FTransUNet [17] combines CNNs and Transformers to hierarchically fuse shallow and deep features for remote sensing image semantic segmentation, effectively capturing both local details and global semantics. LOGCAN++ [18] incorporates global and local category perception modules, effectively addressing the challenge of large scale variations in remote sensing images. RTCNet [19] designed a three-branch architecture for fine details, global context, and crack boundaries, respectively, achieving fast and accurate remote sensing image segmentation.

Despite the significant progress made by these algorithms in the field of semantic segmentation, there are still two key challenges that remain to be addressed. First, effectively modeling global context and local details in complex scenes is still difficult. In cases with high inter-class similarity, existing methods often fail to accurately capture fine-grained structures and object boundaries, leading to semantic confusion, as illustrated in the red box of Figure 1. Second, contextual feature fusion often lacks explicit structural guidance. Many approaches rely on simple stacking or weighted fusion of multi-scale features without dedicated spatial–semantic modeling, which may result in redundant contextual representations.

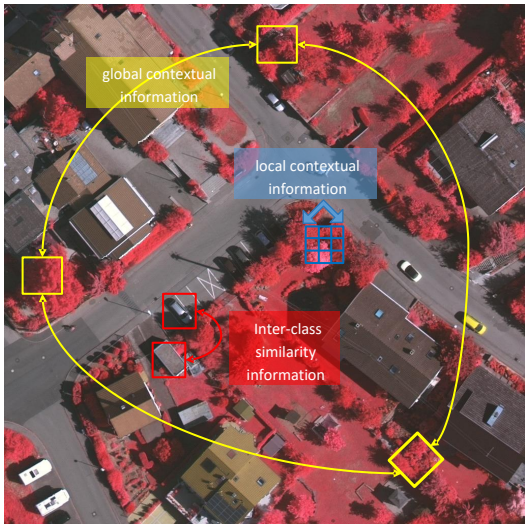

Figure 1. Global–local context information illustration. Blue squares indicate local context information, yellow squares represent intra-class local context information, and red squares represent inter-class local context information. Yellow arrows denote intra-class global context interaction, while red arrows denote inter-class global context interaction.

To address the above issues, this paper proposes an improved semantic segmentation network named CFENet. The main contributions of this work are as follows:

1. We propose a Wavelet-based Pyramid Pooling Module (WPPM) to enhance the model's capability in capturing multiscale targets.
2. We design an Adaptive Context Enhancement (ACE) module to improve the recognition and representation of diverse objects in complex scenes.
3. We introduce a Multilevel Feature Reconstruction (MFR) module to enhance the model's capability of modeling complex structures.
4. We propose a novel deep learning framework, CFENet, which achieves excellent segmentation performance on two public remote sensing datasets.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 provides a detailed description of the proposed algorithm. Section 4 presents experimental results on two public datasets, along with comparisons against baseline and several state-of-the-art models. Sections 5 and 6 are the discussion and conclusion parts of this paper.

2. Related Work

In this section, we briefly review several key concepts relevant to our work, including remote sensing image semantic segmentation, attention mechanism, multiscale convolution, and semantic consistency.

2.1. Remote Sensing Image Semantic Segmentation

High-resolution remote sensing imagery is characterized by complex backgrounds, small inter-class appearance differences, and large intra-class variability. For example, buildings and roads may exhibit highly similar textures under different illumination or sensor conditions, while objects of the same category can look remarkably different across regions. In addition, remote sensing images contain objects with a very wide range of scales—from small individual structures to large building complexes—making it difficult for single-scale representations to capture all types of land-cover features effectively. Moreover, high-resolution scenes often include spatial patterns with strong global structural

dependencies, such as road networks and urban layouts, which impose greater demands on the model's ability to capture long-range contextual relationships.

To address these challenges, researchers have proposed a variety of improved strategies. PAN [20] enhances fine-grained details through pyramid attention, while HRNet [21] maintains high-resolution representations via parallel multiscale branches. Gated-SCNN [22] adopts a dual-branch architecture that separately models geometric boundaries and semantic regions, and UNet3+ [23] employs densely connected cross-layer aggregation to substantially improve multiscale feature fusion. In addition, lightweight models such as BiSeNet [24] and Fast-SCNN [25] leverage the collaboration between shallow spatial branches and deep semantic branches to balance high-resolution detail preservation and real-time performance. EMANet [26] enhance long-range contextual modeling through adaptive context encoding.

UNetFormer integrated lightweight convolutions with hierarchical Transformers, achieving strong performance on remote sensing datasets [27]. RTMamba leverages VSS to extract deep features while maintaining linear computational complexity, thereby capturing long-range contextual information and achieving promising results on remote sensing datasets [28]. RS^{3}Mamba also utilizes VSS to construct an auxiliary branch and introduces a Collaborative Completion Module (CCM) to enhance and fuse features from dual encoders, demonstrating strong segmentation performance on remote sensing data [29]. MSGCNet proposes a multiscale interaction module to bridge the semantic gap between shallow and deep features, and further introduces a Scale-Aware Fusion (SAF) module for efficient decoding [30]. CMTFNet builds a Transformer decoder based on a multiscale multihead self-attention module and designs a multiscale Attention Fusion (MAF) module to fully integrate semantic information from different levels [31].

Although the aforementioned networks have achieved some success in semantic segmentation, limitations remain in jointly modeling global context and local details, as well as efficiently integrating multi-scale feature representations. To address these challenges, we propose a context-aware feature enhancement network, named CFENet, which unifies multi-scale feature extraction, context modeling, and adaptive feature refinement within a single framework to achieve more accurate and robust segmentation.

2.2. Attention Mechanism

In recent years, attention mechanisms have become a fundamental component in semantic segmentation owing to their strong capability to enhance feature representation and contextual modeling. By adaptively reweighting feature responses along spatial or channel dimensions, attention mechanisms enable networks to selectively emphasize informative regions while suppressing irrelevant background noise. This selective feature enhancement is particularly beneficial for complex scenes, where accurate boundary delineation and fine-grained semantic discrimination are critical for high-quality segmentation results.

Early attention-based methods mainly focus on enhancing feature representations through channel-wise or spatial reweighting. SENet [32] exploits global channel statistics via squeeze-and-excitation operations, providing an efficient way to improve feature discriminability; however, the reliance on global pooling inevitably compresses spatial structure information, limiting its ability to capture scale variations and fine-grained contextual cues. CBAM [33] extends this paradigm by sequentially introducing spatial attention, enabling joint modeling of "what" and "where" to attend, yet its attention mechanisms remain locally applied and do not explicitly address cross-scale feature interactions. To better capture long-range contextual dependencies, DANet [34] introduces position and channel attention modules that explicitly model global semantic correlations, significantly enhancing contextual understanding in complex scenes. Nevertheless, such dense attention

mechanisms are typically performed on single-scale feature maps, making them less effective in handling the pronounced scale variations commonly observed in high-resolution remote sensing imagery. CCNet [35] alleviates the computational burden of dense attention by adopting a criss-cross attention strategy, which efficiently aggregates global spatial context with reduced complexity, but its contextual modeling is still confined to a fixed feature resolution.

To address the above limitations, we propose an Adaptive Context Enhancement (ACE) module that explicitly bridges global contextual modeling and local feature refinement. Unlike conventional attention mechanisms that operate on fixed scales or require dense pairwise interactions, ACE adaptively captures contextual dependencies and enhances feature representations in a computationally efficient manner, thereby facilitating multi-scale semantic representation.

2.3. Multiscale Convolution

Standard convolution is a fundamental operation in image processing, capable of extracting local features from input data. CNNs build upon this operation with a hierarchical structure, where different layers have varying receptive fields, providing a basis for multi-scale feature extraction. Repeatedly applying standard convolutions remains insufficient for effectively capturing long-range context or fine-grained boundaries. To overcome these limitations, researchers have developed various multiscale convolution modules, which enhance the network's representational capacity and improve its adaptability to objects of different scales.

The Inception series [36–38] employed parallel convolutional kernels of varying sizes to fuse multiscale information along the spatial dimension. This provided effective local feature extraction, but the fixed kernel sizes limit flexibility in capturing global context and result in relatively high computational cost. DilatedNet [39] introduced dilated (atrous) convolution to expand the receptive field without increasing parameters. It enables richer contextual modeling; however, single-scale dilated convolutions may suffer from gridding artifacts and lack fine-grained feature awareness. DenseASPP [40] further enhanced multiscale perception by densely connecting multiple dilation rates, significantly improving feature representation across scales. This comes at the expense of increased computational and memory cost. PSPNet [41] proposed the Pyramid Pooling Module (PPM) to aggregate multiscale pooled features before convolutional layers, improving global semantic understanding. However, the pooling operations may reduce spatial resolution, limiting fine-grained detail recovery. DAPPM [42] performs deep aggregation of multiscale features, enriching contextual embedding. Its deep architecture and large number of channels per scale, however, hinder parallel computation during inference. PIDNet [43] addresses these efficiency issues by modifying the DAPPM connection strategy and reducing channels per scale. This enables faster parallel context aggregation, though slightly sacrificing context richness.

The aforementioned methods primarily leverage the inherent properties of convolution to capture multiscale features. However, traditional pooling operations, which are often used alongside these convolutions, struggle to adapt to different image regions, making it difficult to flexibly capture targets with drastic scale changes. Moreover, increasing the pooling window size typically leads to substantial information loss. To this end, we design two multiscale convolution modules in this work: WPPM and MFR. WPPM incorporates wavelet-based downsampling [44] to obtain multiscale information while reducing information loss. MFR acquires multiscale features by repeatedly utilizing parallel convolutional branches in an efficient manner.

2.4. Semantic Consistency

Semantic consistency refers to ensuring the semantic alignment and coherence between features from different hierarchical levels when fusing shallow and deep features, thereby avoiding semantic conflicts or ambiguities caused by level discrepancies. In semantic segmentation tasks, shallow features typically contain rich spatial detail information, while deep features embody more abstract semantic information. Due to the mismatch between these feature types, direct fusion often adversely affects the model's prediction accuracy. To address this issue, numerous studies have focused on enhancing semantic consistency among features to bridge the hierarchical gap and improve fusion effectiveness.

Dong et al. proposed SSCCL, which learns semantically and spatially consistent features by maximizing similarity between overlapping regions of augmented views [45]. CMNeXt introduced a lightweight semantic bridging module to enhance fusion robustness and structural expressiveness [46]. SegNeXt proposed a Semantic-Guided Feature Mixer to dynamically adjust shallow encoder features under the guidance of deeper features, improving semantic coherence and boundary awareness [47]. ZMNet employed a multilevel attention feature fusion module and a boundary supervision module to progressively fuse features while mitigating semantic mismatches [48]. SFANet introduced a Stage-aware Feature Alignment module that aligns feature fusion across encoder stages, reducing semantic inconsistencies during decoding [49].

Although the above methods have made meaningful progress in alleviating semantic misalignment during shallow–deep feature fusion, they often overlook the synergy between low-frequency semantic information and high-frequency boundary details. As a result, they struggle to balance semantic coherence and boundary clarity in the fused output. To address this limitation, Chen et al. proposed the Frequency-aware Feature Fusion (FF) framework, which interprets the problem of semantic consistency from a frequency-domain perspective [50]. They pointed out that the instability of feature semantics during fusion essentially arises from the imbalance between different frequency components. Specifically, the high-frequency components within object regions usually correspond to texture variations or noise, and when these signals are directly upsampled without regulation, they introduce fluctuations in intra-class features, thereby degrading semantic consistency. In contrast, the high-frequency components near object boundaries are often attenuated or lost during downsampling and interpolation, resulting in blurred edges and boundary displacement. Then, FreqFusion explicitly distinguishes and adaptively processes different frequency components during feature fusion: it suppresses the redundant high-frequency disturbances within homogeneous regions to enhance the stability of low-frequency semantic information, while simultaneously restoring and strengthening high-frequency details around boundaries to preserve structural sharpness. From the perspective of frequency energy distribution, this method reveals the fundamental cause of semantic inconsistency in feature fusion and achieves a balanced representation between low- and high-frequency information, leading to smoother feature transitions and more accurate semantic boundaries.

3. Method

Considering the complex backgrounds and pronounced inter-class variations characteristic of remote sensing images, we propose CFENet to better accommodate these challenges. The overall architecture of CFENet is illustrated in Figure 2.

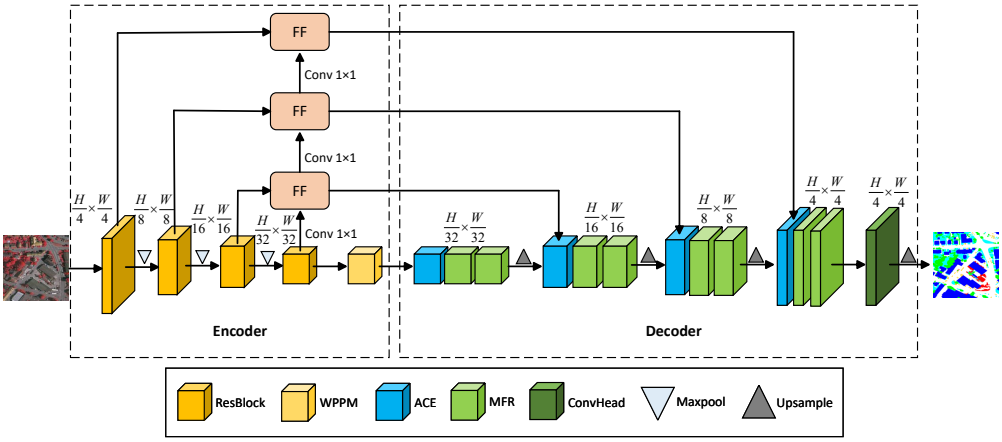

Figure 2. The structure of Context-aware Feature Enhancement Network. Different colored rectangles represent distinct modules in the encoding-decoding process, inverted triangles indicate downsampling, upright triangles indicate upsampling, and FF denotes the Frequency-Aware Feature Fusion module.

Specifically, we adopt a pretrained ResNet18 [51] as the encoder, which consists of four sequential residual blocks, enabling efficient feature encoding with minimal additional parameters and computational cost. In the decoding stage, the WPPM first further processes the output from the encoder, followed by the FF module enhancing semantic consistency between adjacent encoder outputs. The ACE module then performs global–local detailed contextual modeling to achieve feature enhancement. Meanwhile, the MFR module reconstructs fine details of the enhanced features, improving feature representation capability while minimizing feature loss.

3.1. Wavelet-Based Pyramid Pooling Module

In encoder-decoder architectures, the features output by deep encoder layers often contain limited information due to progressive downsampling, which can hinder accurate reconstruction of fine-grained details. To address this, pyramid-based feature extraction has been widely adopted, allowing the network to capture multiscale contextual representations from deep features. However, conventional pooling operations used in these processes inevitably lead to information loss, while repeated convolution operations can introduce redundant features, increasing computational cost without enhancing representational efficiency. To overcome these challenges, we design WPPM. WPPM replaces standard convolution with group convolution, significantly reducing the number of model parameters. In addition, it incorporates haar wavelet downsampling (HWD) [44], which processes and leverages four sub-bands generated during the Haar wavelet transform, enabling the model to retain more information across scales without introducing additional parameters.

The structure of HWD is shown in Figure 3. HWD first applies haar wavelet transform to the input features. H_0 and H_1 denote the low-pass and high-pass decomposition filters, respectively, which are used to extract low-frequency and high-frequency information from the image. The $\downarrow 2$ represents the downsampling applied to both the low-frequency and high-frequency components. Through the Haar wavelet transform, one low-frequency approximation component and three high-frequency detail components (horizontal, vertical, and diagonal directions) are obtained. Their spatial dimensions are reduced by half, while the number of channels in the feature maps increases fourfold. Subsequently, a 1×1 convolution is employed to further reduce the channel count and extract representative features.

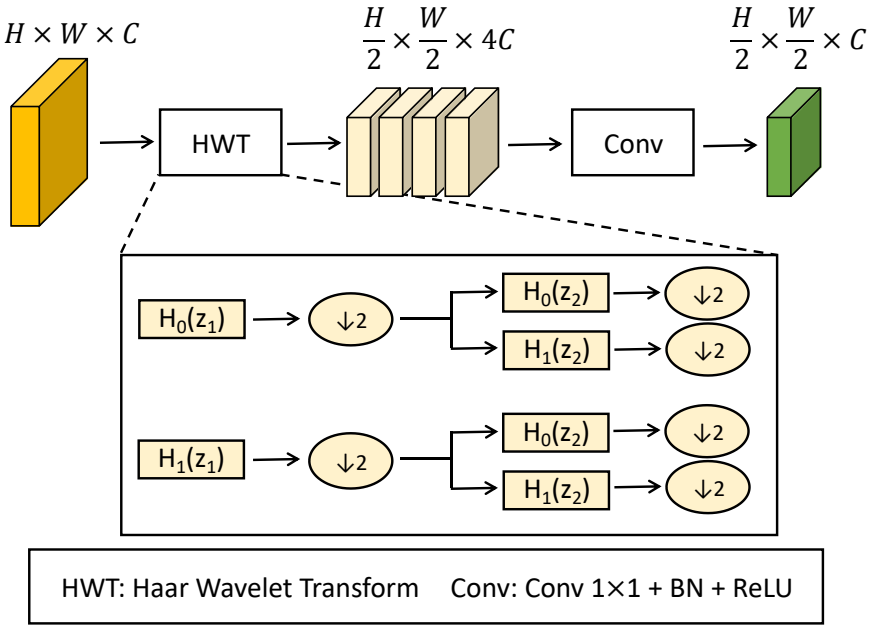

Figure 3. The structure of Haar Wavelet Downsampling Module.

The structure of WPPM is shown in Figure 4. The WPPM first applies a 1×1 convolution followed by normalization to the input features for channel reduction, which helps to decrease the number of parameters and computational cost. Then, the output features are passed through HWD in sequence to capture hierarchical global–local contextual representations. After each wavelet downsampling operation, bilinear interpolation is used to restore the feature maps to the original resolution. These are then fused with the features from the previous level through element-wise addition, achieving spatial alignment of multiscale features while preserving rich detail and semantic information. To further accelerate training, WPPM employs 3×3 group convolution for fine-grained feature reconstruction, and then concatenates it with the first-level output to obtain a preliminary global–local context representation. Finally, a 1×1 convolution is applied to increase the channel dimension of the feature maps, which are then fused with the original input through a residual connection. The process can be expressed by the following equations:

$$X_1 = \text{Conv}\,1 \times 1(X_{in}) \tag{1}$$

$$X_2 = \text{HWD}(X_1) \tag{2}$$

$$X_3 = \text{HWD}(X_2) \tag{3}$$

$$X_4 = \text{HWD}(X_3) \tag{4}$$

Specifically, $X_{in} \in \mathbb{R}^{c \times h \times w}$ denote the input feature map, where c, h and w represent the number of channels, height, and width of the input image. Conv 1×1 denotes a standard 1×1 convolution. HWD represent the applied Haar wavelet downsampling operation. Subsequently, X_2, X_3, and X_4 are upsampled by a factor of 2, 4 and 8 respectively, and then added element-wise with X_1 to obtain Y_2, Y_3 and Y_4. These features are then processed in parallel using group convolution. Finally, a residual connection is applied with the original input X_{in} to produce the output Z.

$$Z = \text{Conv}1 \times 1(X_{in}) + \text{Conv}1 \times 1(\text{Concat}(\text{GroupConv}(Y_2, Y_3, Y_4),\ X_1)) \tag{5}$$

Here, GroupConv denotes group convolution, and Concat represents the concatenation of the output feature maps along the channel dimension.

Through this series of designs, WPPM not only demonstrates stronger capability in modeling complex contextual relationships but also maintains a lightweight structure that is well-suited for efficient integration with backbone networks, significantly enhancing the model's ability to perceive multiscale targets.

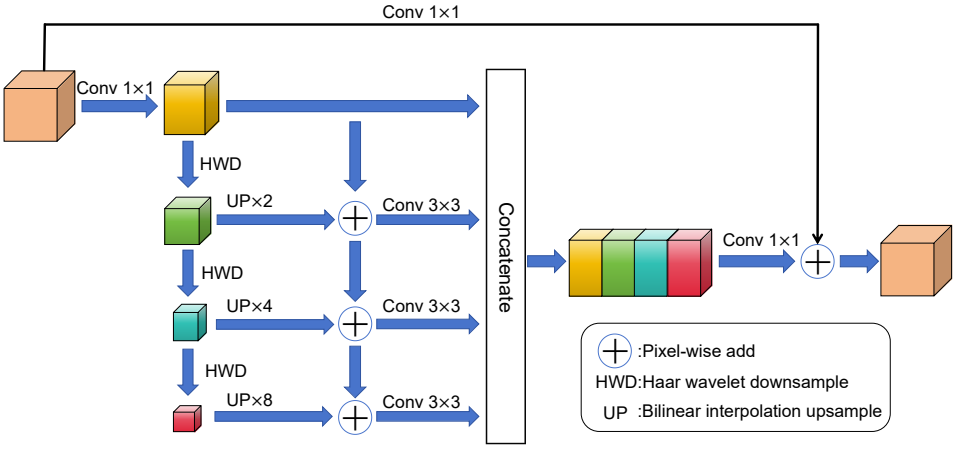

Figure 4. The structure of Wavelet-based Pyramid Pooling Module.

3.2. Frequency-Aware Feature Fusion

The FF module enhances feature fusion by adaptively modulating the frequency components of the features. This module consists of three key components: (1) Adaptive Low-Pass Filter (ALPF): used to smooth intra-class inconsistency in high-level features; (2) Offset Generator: guided by local similarity, it replaces inconsistent features with neighboring consistent representations; (3) Adaptive High-Pass Filter (AHPF): enhances high-frequency details from shallow features to sharpen object boundaries. The module first compresses and preliminarily fuses the high- and low-level features. Then, the ALPF is applied to the high-level features for smoothed upsampling. The Offset Generator performs resampling to further improve feature consistency. Finally, the AHPF extracts high-frequency details from the shallow features and fuses them with the resampled features for final output. The specific structure of FreqFusion is shown in Figure 5.

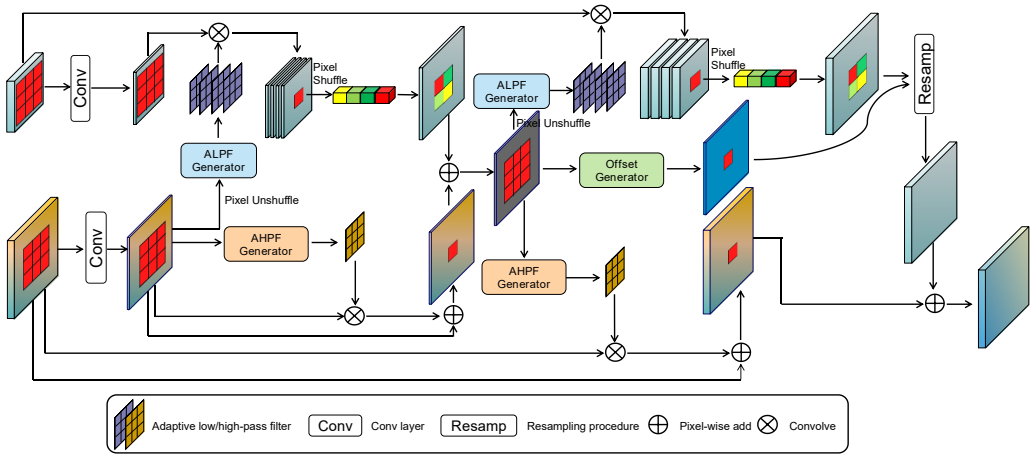

Figure 5. The structure of Frequency-aware Feature Fusion.

In our design, we begin by taking the output of the third encoder layer as the shallow input feature. Since the fourth encoder layer aggregates richer semantic information, it is treated as the deep input feature and fused with the shallow input via FF to obtain semantically enhanced features. This enhanced representation is then used as the new deep

input and fused with the second encoder layer output. The resulting feature is again fused with the first encoder layer output. The complete FF process can be formulated as:

$$Y_3 = f(X_3, X_4) \tag{6}$$

$$Y_2 = f(X_2, Y_3) \tag{7}$$

$$Y_1 = f(X_1, Y_2) \tag{8}$$

Here, $X_i \in \mathbb{R}^{c_i \times h_i \times w_i}$ denotes the output feature map from the encoder, where $i \in \{1, 2, 3, 4\}$ indicates the index of the encoder layer. f denotes the frequency-aware fusion function. $k \in \{1, 2, 3\}$ denotes different levels in the FF module, and $Y_k \in \mathbb{R}^{c_i \times h_i \times w_i}$ represents the enhanced feature obtained through FF. This design effectively reduces semantic inconsistency between encoder outputs and improves the overall quality of feature representations, laying a solid foundation for accurate and efficient decoding in the later stages.

3.3. Adaptive Context Enhancement Module

Current attention mechanisms commonly used in feature enhancement can be broadly categorized into spatial attention and channel attention. Spatial attention effectively highlights informative regions in the feature map, enabling the network to focus on salient spatial structures; however, it often fails to fully suppress irrelevant background information and may overlook fine-grained contextual dependencies. Channel attention, on the other hand, selectively recalibrates feature channels to emphasize important semantic information, but it typically ignores spatial relationships, limiting its ability to preserve detailed structural cues. To achieve more precise and comprehensive context modeling, we propose the Adaptive Context Enhancement (ACE) module. ACE integrates features from shallow and deep layers using a spatially adaptive weighting mechanism, addressing the shortcomings of conventional spatial attention by enhancing relevant regions while maintaining fine-grained details. The module further incorporates a Rectangular Self-Calibration Module (RCM) to model foreground information and suppress cluttered background, improving the discrimination of target regions. Finally, an Efficient Channel Attention (ECA) module is applied to capture channel-wise dependencies, complementing the spatial modeling and reinforcing semantic feature representation.

As illustrated in Figure 6, the ACE module consists of three sequential components: (1) Spatial Attention Fusion Block (SAFB); (2) Rectangular Self-Calibration Module (RCM); (3) Efficient Channel Attention (ECA) module.

Inspired by [43], we design a Spatial Attention Fusion Block to enhance spatial information modeling. SAFB takes as input the output F_{high} from each decoder stage and the corresponding enhanced encoder feature F_{low}. First, F_{high} is upsampled by a factor of 2 to match the spatial resolution of F_{low}. Then, a spatial attention mechanism (SA) [33] is applied to generate attention weights, which are used to independently reweight both F_{high} and F_{low} via element-wise multiplication. The reweighted features are then fused through element-wise addition to obtain an initial fused representation. Next, a standard 3×3 convolution is applied to refine the fused features, enhancing the spatial perception ability of the network. Finally, the refined features are fused again via element-wise addition with F_{low} and the upsampled F_{high} to produce the final output. The output features generated by SAFB are enriched with spatially-aware semantic information. This process can be described by the following equations:

$$F_{up} = \text{Upsample}(F_{high}) \tag{9}$$

$$\alpha = \text{SA}(F_{up}, F_{low}) \tag{10}$$

$$F_{out1} = \text{Conv}\,3 \times 3\big(\alpha \cdot F_{up} + (1-\alpha) \cdot F_{low}\big) + F_{up} + F_{low} \tag{11}$$

where upsample denotes bilinear interpolation upsampling, SA represents spatial attention, and α denotes the weight matrix generated by SA.

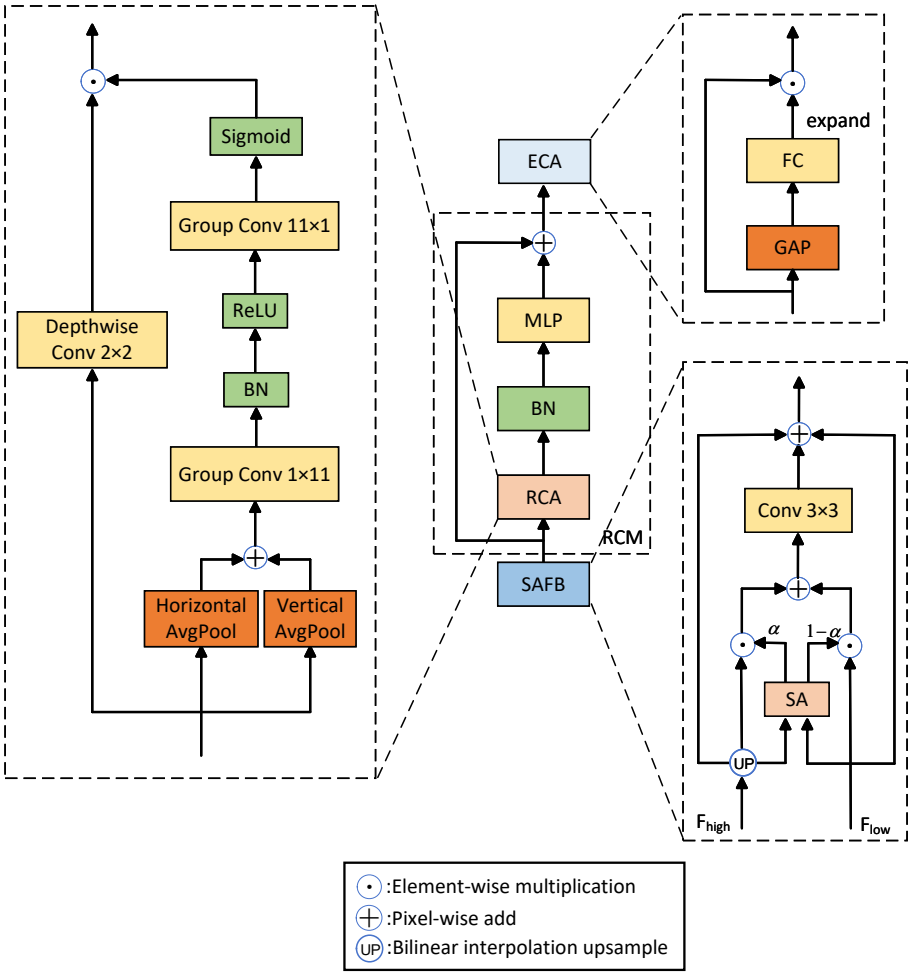

Figure 6. The structure of Adaptive Context Enhancement Module.

The Rectangular Self-Calibration Module is designed to guide the model to focus more on foreground regions within the image [52]. This module introduces axial global contextual information to achieve pyramid-like enhancement of features. It mainly consists of three components: the Rectangular Self-Calibration Attention (RCA), a Batch Normalization (BN) layer, and a Multi-Layer Perceptron (MLP). In the RCA, the module performs global pooling along horizontal and vertical directions, respectively, to obtain two independent axial context vectors. These vectors are then fused via broadcast addition, effectively emphasizing structural information within rectangular regions. Next, direction-aware large strip convolutions are applied separately to decouple and calibrate the attention map: first, a horizontal strip convolution adjusts feature distribution along the row direction to better align with object contours. After BatchNorm and ReLU activation, a vertical strip convolution is applied along the column direction to further enhance vertical structural representation. Subsequently, RCA applies a depthwise separable convolution to extract fine-grained details from the input feature map. The calibrated attention weights

are then fused with the original features via element-wise multiplication, enhancing the discriminative power of the feature representations. At the end of the module, Batch Normalization and an MLP are sequentially applied to further refine the features, followed by a residual connection to facilitate feature reuse. This architectural design improves the model's ability to adapt to targets of various shapes and scales, making it particularly effective for foreground modeling in complex scenes. This process can be described by the following equations:

$$F_1 = \mathrm{HAvg}(F_{out1}) + \mathrm{VAvg}(F_{out1}) \tag{12}$$

$$\beta = \mathrm{Sigmoid}(\mathrm{GroupConv}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{GroupConv}(F_1))))) \tag{13}$$

$$\mathrm{RCA} = \beta \cdot \mathrm{DWConv}(F_{out1}) \tag{14}$$

$$F_{out2} = \mathrm{MLP}(\mathrm{BN}(\mathrm{RCA}(F_{out1}))) + F_{out1} \tag{15}$$

here, HAvg and VAvg denote horizontal average pooling and vertical average pooling, respectively. The BN stands for batch normalization, ReLU and Sigmoid are activation functions, DWConv refers to depthwise convolution, and GroupConv represents group convolution.

The Efficient Channel Attention module employs a local cross-channel interaction strategy to achieve fast channel-wise attention allocation, significantly enhancing semantic feature discrimination while maintaining low computational overhead [53]. Specifically, after global average pooling, a lightweight one-dimensional convolution with kernel size k is applied to enable local cross-channel information exchange. The kernel size k controls the coverage of local cross-channel interactions, indicating how many neighboring channels participate in predicting the attention weight for each channel. Finally, the generated attention weights are applied to the original input via element-wise multiplication.

Through the coordinated integration of these three components, the ACE module achieves efficient context modeling and enhances the model's recognition capability in complex scenes.

3.4. Multiscale Feature Reconstruction Module

Deep features extracted by the encoder contain rich semantic and structural information, which is crucial for accurate reconstruction during decoding. Convolutional operations are naturally suited for this task, as they can effectively integrate local context and progressively refine feature representations. However, relying solely on standard convolutional layers often limits the ability to capture complex edges and fine-grained boundary information, particularly in high-resolution or structurally intricate scenes. To address this limitation, we propose the Multiscale Feature Reconstruction (MFR) module, which adopts a parallel pure-convolution architecture capable of adapting to input features of different scales. This design allows the module to retain more information from deep features, facilitating the reconstruction of complex patterns and subtle edges. Furthermore, MFR employs element-wise multiplication to project features into a high-dimensional implicit feature space, enhancing the network's capacity to model subtle edges and intricate patterns that standard convolution alone may overlook. Compared with existing multiscale reconstruction modules, MFR's combination of parallel convolutional processing and high-dimensional feature mapping enables more flexible and precise multiscale feature reconstruction, improving boundary refinement and overall segmentation performance. The specific structure of MFR is shown in Figure 7.

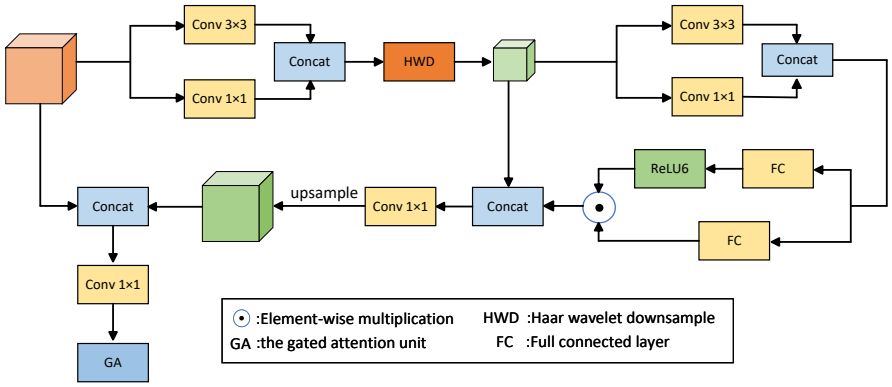

Figure 7. The structure of Multiscale Feature Reconstruction Module.

In Figure 7, MFR begins with a dual-branch design that applies both 3×3 and 1×1 convolutions to the input features. The outputs of these two branches are then concatenated along the channel dimension and passed through HWD to perform spatial compression and enhance local feature representation. Subsequently, a feature integration unit (a 1×1 convolution) is applied to compress the channel dimensions and aggregate semantic information, producing a compact initial semantic feature map, whose spatial size is half that of the input. To ensure the effective utilization of semantic information, the output features from the previous layer are once again processed through the same dual-branch structure (3×3 and 1×1 convolutions). Meanwhile, the study of StarNet [54] demonstrates that high-dimensional spaces can offer greater information capacity and expressive power, as they are capable of capturing subtle differences and complex patterns among data. Element-wise multiplication, in particular, has the ability to project features into a high-dimensional implicit feature space. Therefore, we adopt a 1×1 convolution instead of a fully connected layer to perform a linear transformation, followed by ReLU6 activation. The transformed features are then projected into a high-dimensional implicit space through element-wise multiplication, resulting in more expressive feature representations. This information is concatenated with the initially refined semantic features and further fused using an information integration unit (1×1 convolution) to generate the final deeply refined semantic features. Since the feature maps are downsampled during the refinement process, we apply bilinear interpolation to upsample the deeply refined semantic features. These are then concatenated with the original input once again. As a result, the output from the information integration unit incorporates refined semantic information from multiple scales and possesses enhanced representational capacity. Additionally, a gated attention unit is introduced to establish a cross-scale residual fusion path between the reconstructed features and the original input, further enhancing the complementary relationship among multiscale features [55]. Through hierarchical reconstruction, feature interaction, and channel-wise selection, the MFR module significantly improves the discriminative capability and multi-scale adaptability of the network while maintaining low redundancy. This enhances the model's ability to capture complex object structures.

$$X_1 = \mathrm{ReLU6}\left(\mathrm{BN}\left(\mathrm{HWD}(\mathrm{Concat}(\mathrm{Conv}\,3\times3(X_{in}), \mathrm{Conv}\,1\times1(X_{in})))\right)\right) \quad (16)$$

$$X_2 = \mathrm{Concat}(\mathrm{Conv}\,3\times3(\mathrm{Conv}\,1\times1(X_1)), \mathrm{Conv}\,1\times1(\mathrm{Conv}\,1\times1(X_1))) \quad (17)$$

$$X_3 = \mathrm{ReLU6}(\mathrm{FC}_1(X_2)) \cdot \mathrm{FC}_2(X_2) \quad (18)$$

$$X_4 = \mathrm{Concat}(\mathrm{Upsample}(\mathrm{Conv}\,1\times1(\mathrm{Concat}(X_1, X_3))), X_{in}) \quad (19)$$

$$X_{out} = \mathrm{ReLU6}(\mathrm{BN}(\mathrm{GA}(\mathrm{Conv}\,1\times1(X_4)))) \quad (20)$$

where BN denotes batch normalization, ReLU6 represents the activation function, FC_1 and FC_2 refer to the two fully connected layers, and GA indicates the gated attention unit.

4. Experiments and Results

In this section, we first introduce the experimental settings, evaluation metrics, loss functions, and the datasets used. We then validate the effectiveness of the proposed modules on the Vaihingen and Potsdam datasets. Finally, we perform both quantitative and qualitative comparisons between our method and other mainstream approaches.

4.1. Experimental Environment

All experiments were conducted on a Linux system with Python 3.10.15, PyTorch 2.4.1, and CUDA 11.8. Model training was performed on a single NVIDIA A30 GPU with 24 GB of memory. The input image resolution was set to 1024 × 1024 pixels. During training, various data augmentation strategies were employed, including random scaling (with scale factors of [0.5, 0.75, 1.0, 1.25, 1.5]), random vertical flipping, random horizontal flipping, and random rotation. The batch size was set to 4, and the optimizer used was AdamW with a learning rate of 0.0006. The number of training epochs was set to 100. During testing, multiscale and random flipping strategies were also applied to enhance performance. The experimental environment and parameters are summarized in Tables 1 and 2 below.

Table 1. Details of the experimental environment configuration.

Name	Details
Operating system	Linux 4.18
GPU	A30
CUDA	11.8
CUDNN	9.1.0
Python	3.10.15
PyTorch	2.4.1

Table 2. Details of the hyperparameters configured for the experiments.

Full Name	Parameter
Image size	1024 × 1024 × 3
Epochs	100
Batch size	4
Number of workers	4
Learning Rate	0.0006
Weight decay	0.01

4.2. Evaluation Indicator

In this study, we adopt three primary evaluation metrics: (1) Overall Accuracy (OA), (2) Mean Intersection over Union (mIoU), (3) Mean F1 Score (mF1). The main formulas are defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{23}$$

$$OA = \frac{TP}{TP + TN + FN + FP} \tag{24}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{25}$$

where TP, TN, FN, and FP denote true positives, true negatives, false negatives, and false positives, respectively.

4.3. Loss Function

The loss function used in this work is defined as follows:

$$L = L_{ce} + L_{dice} \tag{26}$$

where L_{ce} denotes the cross-entropy loss and L_{dice} denotes the Dice loss. The specific formulas are as follows:

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)} \log\left(p_i^{(c)}\right) \tag{27}$$

$$L_{dice} = 1 - \frac{1}{C}\Sigma_{c=1}^{C}\frac{2\sum_{i=1}^{N} p_i^{(c)} y_i^{(c)}}{\sum_{i=1}^{N} p_i^{(c)} + \sum_{i=1}^{N} y_i^{(c)} + \epsilon} \tag{28}$$

here, N denotes the total number of pixels in the image, and C is the total number of classes. $p_i^{(c)}$ represents the predicted probability that the i-th pixel belongs to class c, while $y_i^{(c)}$ denotes the one-hot encoded ground truth of the i-th pixel for class c. ϵ is a small constant added to avoid division by zero.

4.4. Experimental Datasets

We employed two distinct remote sensing datasets, ISPRS Vaihingen and ISPRS Potsdam, to evaluate the effectiveness of our model.

(1) Vaihingen: This dataset contains high-resolution aerial images captured over the Vaihingen City in Germany, with fine-grained manual annotations for six land cover classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter. The dataset consists of 33 orthorectified images with a spatial resolution of 0.09 m per pixel, and image widths ranging from 1887 to 3816 pixels. For evaluation, 17 images were used for training, 1 for validation, and the remaining 15 for testing. All Top images were cropped into 1024×1024 pixel patches.

(2) Potsdam: This dataset includes 38 high-resolution Top image tiles, each with a size of 6000×6000 pixels, covering the same six land cover classes as the Vaihingen dataset. It provides four multispectral bands, along with auxiliary data such as DSM and NDSM. In our experiments, only the RGB bands were used. A total of 22 images were selected for training, 1 image for validation, and the remaining 14 for testing. The images numbered 7_10 was removed due to annotation errors. The original image tiles were cropped into patches of 1024×1024 pixels. Due to the extremely small number of pixels belonging to the clutter category in the ISPRS Vaihingen and Potsdam datasets, this class is not considered in the quantitative evaluation. This choice helps avoid potential bias in the performance metrics and allows a more accurate assessment of the model's effectiveness on the major semantic classes.

4.5. Ablation Study

In this section, we conduct ablation studies on the overall architecture using the Vaihingen and Potsdam datasets to verify the effectiveness of each proposed improvement, including the introduction of FF and the design of the ACE, MFR, and WPPM. The overall ablation study results are presented in Tables 3 and 4. Additional ablation experiments are shown in Tables 5–9.

Table 3. Ablation study of the new model on the Vaihingen dataset.

Dataset	MFR	ACE	FF	WPPM	mF1 (%)	OA (%)	mIoU (%)
Vaihingen					88.73	92.07	80.33
	✓				90.11	92.71	82.39
	✓	✓			90.84	93.12	83.44
	✓	✓	✓		91.47	93.42	84.29
	✓	✓	✓	✓	91.89	93.61	84.77

Note: ✓ indicates that the corresponding module is included in the model configuration.

Table 4. Ablation study of the new model on the Potsdam dataset.

Dataset	MFR	ACE	FF	WPPM	mF1 (%)	OA (%)	mIoU (%)
Potsdam					90.42	88.93	82.71
	✓				91.35	90.29	84.05
	✓	✓			91.92	90.60	85.13
	✓	✓	✓		92.44	91.34	86.22
	✓	✓	✓	✓	92.80	91.73	86.62

Note: ✓ indicates that the corresponding module is included in the model configuration.

(1) Ablation Study on the Overall Architecture

Table 3 presents the ablation study results of each module on the Vaihingen dataset. We observe that the MFR module to the baseline U-Net network with a pre-trained ResNet-18 backbone enables more efficient feature representation, improving mF1, OA, and mIoU by 1.38%, 0.64%, and 2.03%, respectively. Subsequently, the adding of the ACE module further enhances the network, increasing mF1, OA, and mIoU by 0.73%, 0.41%, and 1.05%, respectively, owing to ACE's effective context modeling. The FF module improves the semantic consistency of encoder outputs, resulting in increments of 0.63%, 0.30%, and 0.85% in mF1, OA, and mIoU. Finally, incorporating the WPPM boosts the model's multiscale modeling capability, raising mF1, OA, and mIoU to 91.89%, 93.61%, and 84.77%, respectively. These results demonstrate the effectiveness of the designed and introduced modules.

Table 4 presents the ablation study results of each module on the Potsdam dataset. Specifically, compared with the baseline network, our model achieves increases of 1.38%, 2.80%, and 3.91% in mF1, OA, and mIoU, respectively, reaching final values of 92.80%, 91.73%, and 86.62%.

(2) Ablation Study on Different Encoders

Table 5 presents an ablation study on different encoder backbones for CFENet on the Vaihingen dataset. As shown, employing stronger encoders such as ResNet50 [51] and Swin-Tiny [56] leads to only marginal improvements in mIoU compared with ResNet18. Specifically, Swin-Tiny achieves the highest mIoU of 85.03%, which is only 0.26% higher than that of ResNet18. However, this minor performance gain comes at the cost of a significantly increased number of parameters and computational complexity. Considering the favorable trade-off between segmentation accuracy and model efficiency, ResNet18 is adopted as the default encoder in the following experiments.

Table 5. Ablation study of different encoders on the Vaihingen dataset.

Method	Backbone	Params (M)	FLOPs (G)	mIoU (%)
CFENet	ResNet50	41.48	202.61	84.79
	Swin-Tiny	44.62	209.28	85.03
	ResNet18	24.68	161.08	84.77

(3) Ablation Study on the MFR Module

We conducted additional ablation experiments on the MFR module using the Vaihingen dataset. We added only the MFR module to the baseline model and varied the kernel

sizes (excluding the 1×1 branch) in the parallel branches. The experimental results are shown in Table 6.

Table 6. Ablation study of convolution kernel sizes in the MFR module on the Vaihingen dataset.

Kernal_Size	mF1 (%)	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
3×3	90.11	92.71	82.39	1.07	6.66
5×5	90.18	92.77	82.42	2.01	11.90
7×7	90.21	92.79	82.46	3.42	19.76

In Table 6, when using 3×3 convolutions, the network achieved an mF1 of 90.11%, OA of 92.71%, and mIoU of 82.39%. But when using 5×5 convolutions with a larger receptive field, there was no significant improvement in performance. On the contrary, the number of parameters and computational cost increased substantially. We preliminarily attribute this to the relatively small size of feature maps in the shallow decoder, where larger kernels do not offer a clear advantage in capturing fine details compared to smaller ones. Further increasing the kernel size to 7×7 still did not improve performance and instead added additional computational burden, further validating our hypothesis. Therefore, we chose to use 3×3 convolutions in the MFR module.

We also conducted an ablation study on the number of stacked MFR modules (denoted as K) in the decoder, and the results are presented in Table 7.

Table 7. Ablation study on the number of stacked MFR modules (K) in the decoder on the Vaihingen dataset.

K	mF1 (%)	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
1	89.12	92.33	81.08	0.63	3.69
2	90.11	92.71	82.39	1.07	6.66
3	90.17	92.72	82.44	1.52	9.62

In Table 7, we found that when $K = 1$, the output from the ACE module is passed through a single MFR block for feature reconstruction, achieving mF1, OA, and mIoU scores of 89.12%, 92.33%, and 81.08%. When $K = 2$, the performance improved by 0.99%, 0.38%, and 1.31% respectively, indicating a noticeable gain; however, this also increased the number of parameters by 2.97 M and the computational cost by 0.44 G. When $K = 3$, the performance did not show significant improvement, while the computational burden further increased. Therefore, we choose to stack the MFR module twice in series in the decoder.

(4) Ablation Study on the WPPM

To verify the rationality of the WPPM design, we conducted additional ablation experiments on the Vaihingen dataset. We compared its performance with the original PAPPM module based on the same baseline model, and the results are shown in Table 8.

Table 8. Ablation experiment of WPPM on the Vaihingen dataset. The numbers of input and output channels were kept consistent throughout the experiment.

Method	mF1 (%)	OA (%)	mIoU (%)
Baseline	88.73	92.07	80.33
Baseline + PAPPM	89.37	92.21	81.05
Baseline + WPPM	89.88	92.38	81.46

As presented in Table 8, incorporating PAPPM improved the baseline model's mF1 score, OA, and mIoU by 0.64%, 0.14%, and 0.72%, respectively. In contrast, incorporating WPPM led to improvements of 1.15%, 0.31%, and 1.13% in mF1, OA, and mIoU, respectively.

We further visualize the effectiveness of the WPPM, as shown in Figure 8. Compared with PAPPM, the proposed WPPM produces more accurate segmentation results for trees and buildings, which can be mainly attributed to the wavelet-based downsampling strategy adopted in the feature extraction stage. Unlike the average pooling operations with multiple window sizes used in PAPPM, the successive wavelet downsampling in WPPM effectively reduces information loss without introducing additional computational overhead. Overall, WPPM leads to more effective feature extraction and consequently delivers more accurate and stable segmentation results.

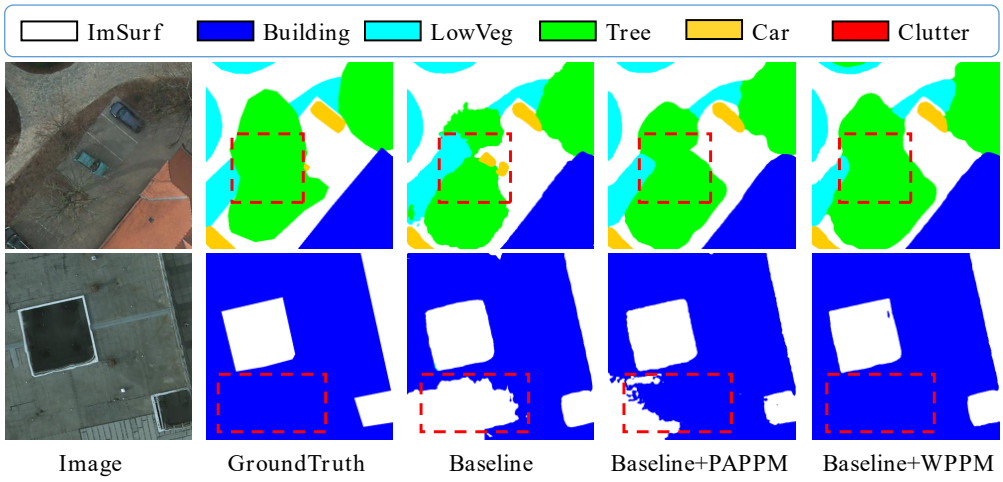

Figure 8. Qualitative comparison of the proposed WPPM with the baseline model and PAPPM on the Vaihingen test set. The red dashed box indicates the parts that require special attention.

(5) Ablation Study on Network Stability

To evaluate the stability of the proposed network, we trained CFENet with different input sizes, including 512 × 512, 1024 × 1024, and 2048 × 2048. The experimental results are shown in Table 9.

Table 9. Ablation Study on different input sizes on the Vaihingen dataset.

Input Size	F1-Score (%)					mIoU (%)
	Imp.	Building	Low.	Tree	Car	
512 × 512	95.28	97.09	85.42	90.10	89.61	84.03
1024 × 1024	95.96	97.31	86.00	90.48	89.71	84.77
2048 × 2048	95.73	97.22	86.01	90.32	88.69	84.25

From Table 9, it can be demonstrated that our network exhibits good stability when processing images of varying resolutions, with only minor variations in mIoU. The model achieves the best performance with an input size of 1024 × 1024. It is worth noting that when the input size increases to 2048 × 2048, small objects such as cars become further reduced in scale, leading to a significant drop in recognition accuracy.

4.6. Comparative Experiments and Results

We compared our model with several mainstream methods on the Vaihingen and Potsdam datasets. The selected methods mainly include convolution-based networks (U-Net [11], PSPNet [41], DeepLabV3+ [57], ABCNet [58] and LOGCAN++ [18]), attention-based networks (MANet [59] and MSGCNet [30]), transformer-based approaches (UNet-

Former [27], DC-Swin [16], CMTFNet [31], SAM2Former [60], BEMS-UNetFormer [61] and DeepKANSeg [62]), and Mamba-based method (RS^{3}Mamba [29]).

(1) Comparison with other methods on the Vaihingen test set

The comparison results with other algorithms on the Vaihingen set are presented in Table 10.

Table 10. Quantitative comparison with state-of-the-art methods on the Vaihingen test set.

Method	Backbone	F1-Score (%)					mF1 (%)	OA (%)	mIoU (%)
		Imp.	Building	Low.	Tree	Car			
PSPNet [41]	ResNet50	90.36	96.11	83.29	89.11	81.77	88.78	92.15	80.27
DeepLabV3+ [57]	-	87.49	93.80	77.41	86.55	61.22	81.30	88.30	69.93
U-Net [11]	-	92.62	95.80	81.84	88.38	80.01	87.73	91.41	78.67
MANet [59]	ResNet50	94.88	96.68	84.09	89.51	85.89	90.21	92.86	82.53
ABCNet [58]	ResNet18	95.48	96.60	85.15	90.46	88.09	91.15	93.30	84.04
UNetFormer [27]	ResNet18	95.16	96.77	84.06	89.99	87.61	90.72	93.04	83.35
DC-Swin [16]	Swin-T	95.56	96.70	84.58	90.31	88.30	91.09	93.24	83.95
RS^{3}Mamba [29]	ResNet18	95.75	96.86	84.55	90.27	88.14	91.11	93.36	84.01
MSGCNet [30]	ResNet18	95.74	97.03	85.45	90.26	89.01	91.50	93.56	84.62
CMTFNet [31]	ResNet50	95.49	96.80	84.95	90.11	89.08	91.29	93.25	84.26
SAM2Former [60]	R18-SAM2	-	-	-	-	-	91.49	92.64	84.56
BEMS-UNetFormer [61]	ResNet18	94.96	96.00	85.36	86.50	86.50	90.59	92.09	83.10
DeepKANSeg [62]	R18-ViT-L	93.46	97.03	82.93	**92.09**	**90.18**	91.14	-	84.05
LOGCAN++ [18]	ResNet50	95.28	96.79	84.42	90.15	89.35	91.20	93.23	84.13
CFENet (ours)	ResNet18	**95.96**	**97.31**	**86.00**	90.48	89.71	**91.89**	**93.61**	**84.77**

Table 10 shows that compared to earlier classic models such as U-Net [11] and PSPNet [41], our model improved mF1 by 4.16% and 3.11%, OA by 2.20% and 1.46%, and mIoU by 6.10% and 4.50%. Meanwhile, compared to some recent mainstream models, such as SAM2Former [60] and BEMS-UNetFormer [61], our model's mF1 increased by 0.40% and 1.30%, OA increased by 0.97% and 1.52%, and mIoU increased by 0.21% and 1.67%, respectively, demonstrating the effectiveness of our approach. Although DeepKANSeg [62] attains slightly higher accuracy on the Tree and Car classes, this improvement can be mainly attributed to its large-scale ViT-L backbone and the nonlinear feature refinement ability introduced by KAN-based modules, which are particularly effective in handling fine-grained or irregular object boundaries. In contrast, our model achieves a more balanced segmentation performance across all land-cover categories, making it more suitable for practical remote sensing applications.

Figure 9 presents a radar-chart-based visualization of the per-class IoU comparisons between the proposed CFENet and four representative state-of-the-art methods on Vaihingen test set. As illustrated in the figure, CFENet achieves consistently competitive segmentation performance across all categories.

(2) Comparison with other methods on the Potsdam test set

The comparison results with other algorithms on the Potsdam test set are presented in Table 11.

Table 11. Quantitative comparison with state-of-the-art methods on the Potsdam test set.

Method	Backbone	F1-Score (%)					mF1 (%)	OA (%)	mIoU (%)
		Imp.	Building	Low.	Tree	Car			
PSPNet [41]	ResNet50	92.00	94.06	85.79	86.03	93.91	90.36	88.96	82.62
DeepLabV3+ [57]	-	88.10	89.92	78.56	75.71	87.76	84.01	82.87	72.84
U-Net [11]	-	89.88	91.08	80.61	78.79	93.37	86.75	84.76	77.07
MANet [59]	ResNet50	92.97	95.62	86.82	87.89	95.65	91.79	90.47	85.05
ABCNet [58]	ResNet18	93.52	95.89	87.27	88.34	95.72	92.15	90.98	85.66
UNetFormer [27]	ResNet18	93.52	96.07	87.01	87.95	95.84	92.08	90.83	86.10
DC-Swin [16]	Swin-T	**94.10**	96.43	87.96	89.06	95.95	92.70	91.52	86.59
RS^{3}Mamba [29]	ResNet18	93.91	96.33	87.40	88.67	96.26	92.52	91.29	86.30
MSGCNet [30]	ResNet18	93.58	96.19	87.88	88.75	96.03	92.49	91.27	86.23
CMTFNet [31]	ResNet50	92.84	95.41	87.34	88.18	95.61	91.88	90.62	85.17
SAM2Former [60]	R18-SAM2	-	-	-	-	-	92.36	91.06	86.01
BEMS-UNetFormer [61]	ResNet18	81.91	94.57	87.93	88.48	95.57	92.43	91.56	86.12
DeepKANSeg [62]	R18-ViT-L	92.69	**97.11**	86.84	86.89	96.30	91.97	-	85.44
LOGCAN++ [18]	ResNet50	92.89	95.63	87.39	88.62	95.72	92.05	90.63	85.46
CFENet (ours)	ResNet18	93.59	96.42	**88.31**	**89.36**	**96.33**	**92.80**	**91.73**	**86.62**

Table 11 shows that compared to early classic models such as U-Net [11] and PSPNet [41], our method improved mF1 by 6.05% and 2.44%, OA by 6.97% and 2.77%, and mIoU by 9.55% and 4.00%. Meanwhile, compared with some recent mainstream models like SAM2Former [60] and BEMS-UNetFormer [61], our model increased mF1 by 0.44% and 0.37%, OA by 0.67% and 0.17%, and mIoU by 0.61% and 0.50%, further demonstrating the effectiveness of our approach. In distinguishing between the two easily confused categories, "LowVeg" and "Tree", our model achieved the best performance, with F1 scores of 88.31% and 89.36%, respectively. Additionally, in the segmentation of small targets, our model achieved an F1 score of 96.33% on the "Car" category, outperforming all comparison methods, indicating excellent performance in small-scale object recognition tasks as well. The consistent improvement across various land-cover categories highlights the effectiveness of our proposed architecture and confirms its strong potential for advancing remote sensing semantic segmentation.

Figure 10 illustrates the per-class IoU comparison between CFENet and four mainstream methods on Potsdam test set using a radar chart. The results indicate that CFENet maintains strong and balanced segmentation performance across different semantic categories.

(3) Qualitative Visualization Analysis

To demonstrate the superior performance of CFENet, we visualized the segmentation results on the Vaihingen test set. The results are shown in Figure 11.

From Figure 11, for the first input image, UNetFormer [27] and RS^{3}Mamba [29] perform poorly in predicting the building category, whereas our proposed model successfully segments the buildings completely. For the second image, CFENet accurately distinguishes impervious surfaces, while other methods, such as MSGCNet [30] and RS^{3}Mamba [29], misclassify these regions as low vegetation due to their similar feature representations. These results further demonstrate that CFENet possesses stronger discriminative capability in semantic recognition.

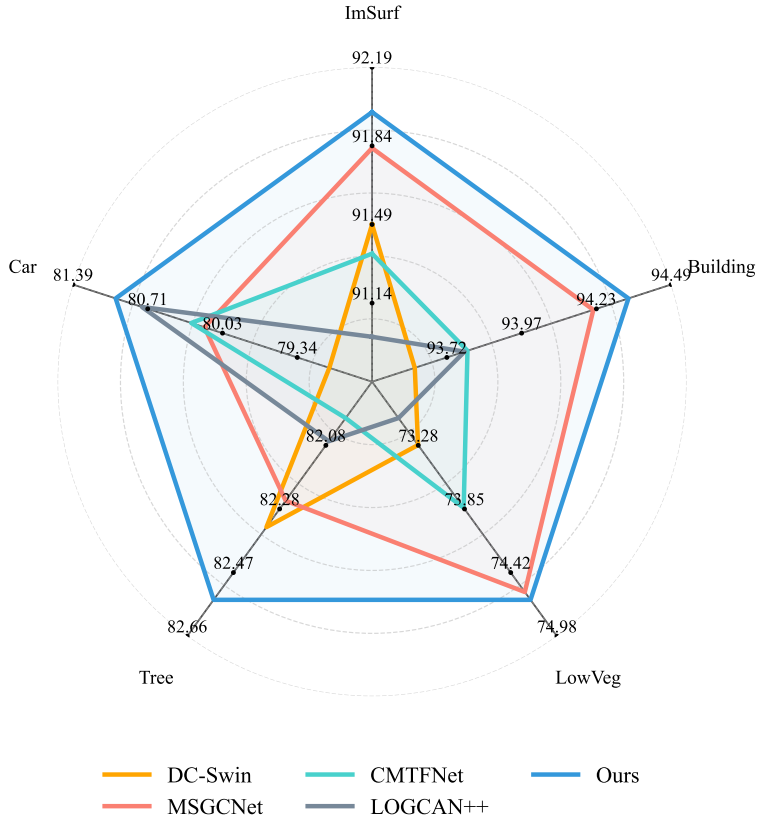

Figure 9. Comparison of IoU for five classes between the proposed CFENet and other models on the Vaihingen test set.

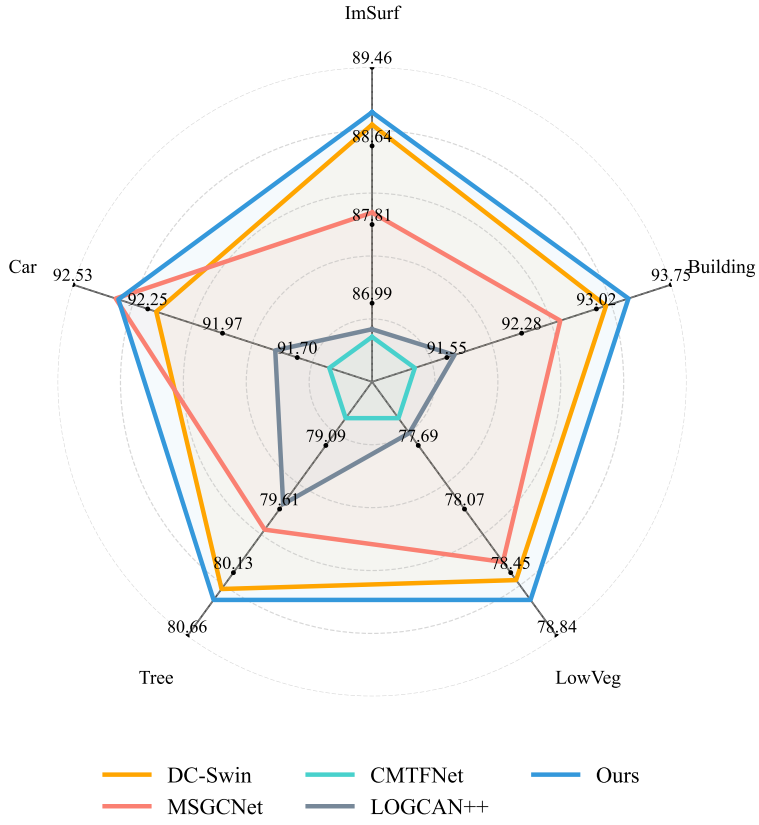

Figure 10. Comparison of IoU for five classes between the proposed CFENet and other models on the Potsdam test set.

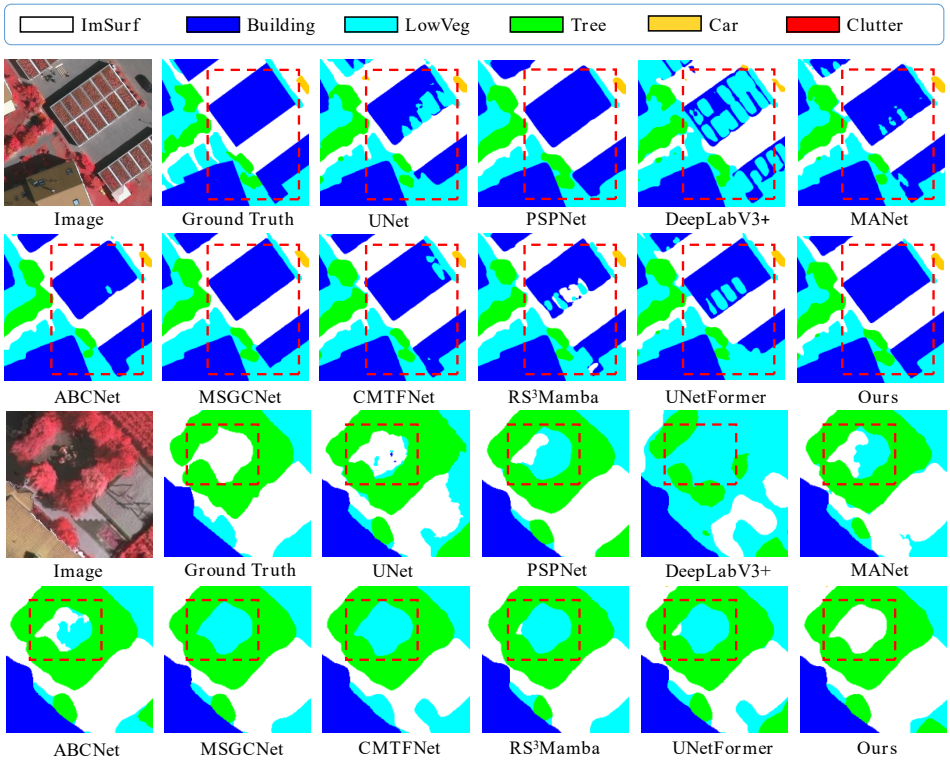

Figure 11. Qualitative comparison of the proposed CFENet with other mainstream models on the Vaihingen test set. The red dotted box indicates the part that needs special attention.

We also visualized the segmentation results on the Potsdam test set. The results are shown in Figure 12. For the first input image, it can be seen that our algorithm improves sensitivity to complex backgrounds, significantly reducing the misclassification rate of the background. In the second input image, the proposed CFENet accurately identifies impervious surfaces, whereas other methods, such as DeepLabV3+ [57] and RS^{3}Mamba [29], misclassify these areas as buildings and background, respectively, further indicating that our algorithm demonstrates strong segmentation performance under complex backgrounds and high inter-class similarity scenarios.

Meanwhile, we present heatmaps of segmentation results for different classes on the Vaihingen test set, comparing the results before and after using the ACE module. The results are shown in Figure 13. Before applying ACE, the model exhibits certain limitations in modeling different categories. Specifically, it struggles to accurately recognize large objects such as buildings, resulting in confusion and blurry segmentation boundaries. Moreover, the recognition of small objects like cars is also imprecise. After incorporating ACE, the model performs more refined contextual modeling for different categories, particularly for buildings and trees, which exhibit significant scale variations. At the same time, the recognition accuracy of small objects such as cars is noticeably improved.

(4) Model Parameters and Computational Complexity Analysis

We use images of size $3 \times 1024 \times 1024$ as input to evaluate the parameter counts and computational complexity of different models. The results are shown in Table 12.

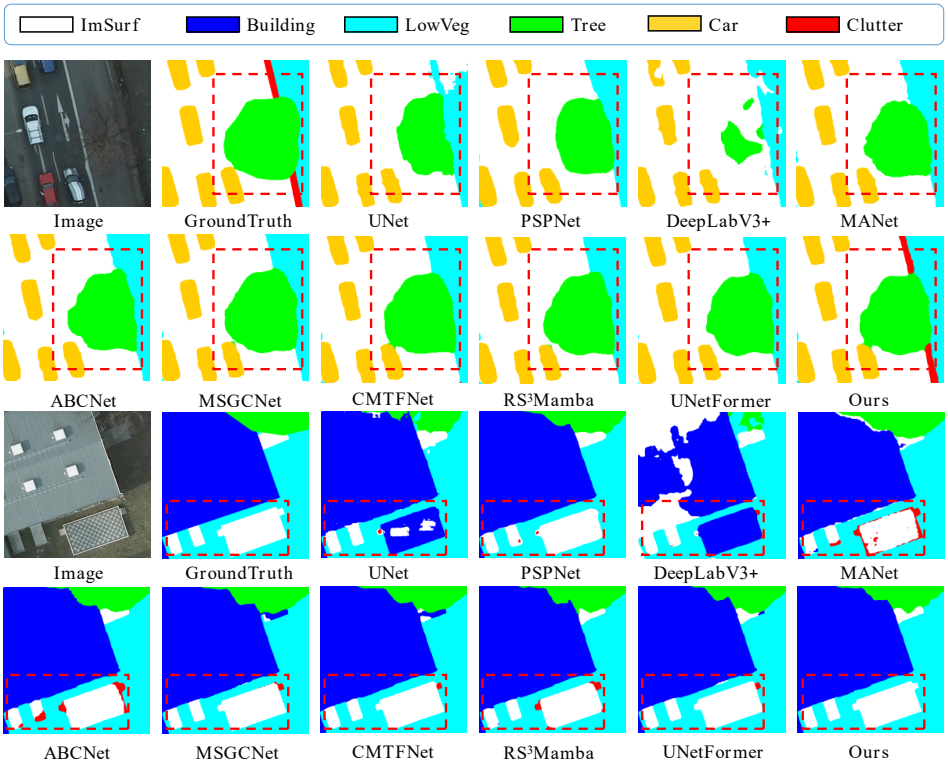

Figure 12. Qualitative comparison of the proposed CFENet with other mainstream models on the Potsdam test set. The red dotted box indicates the part that needs special attention.

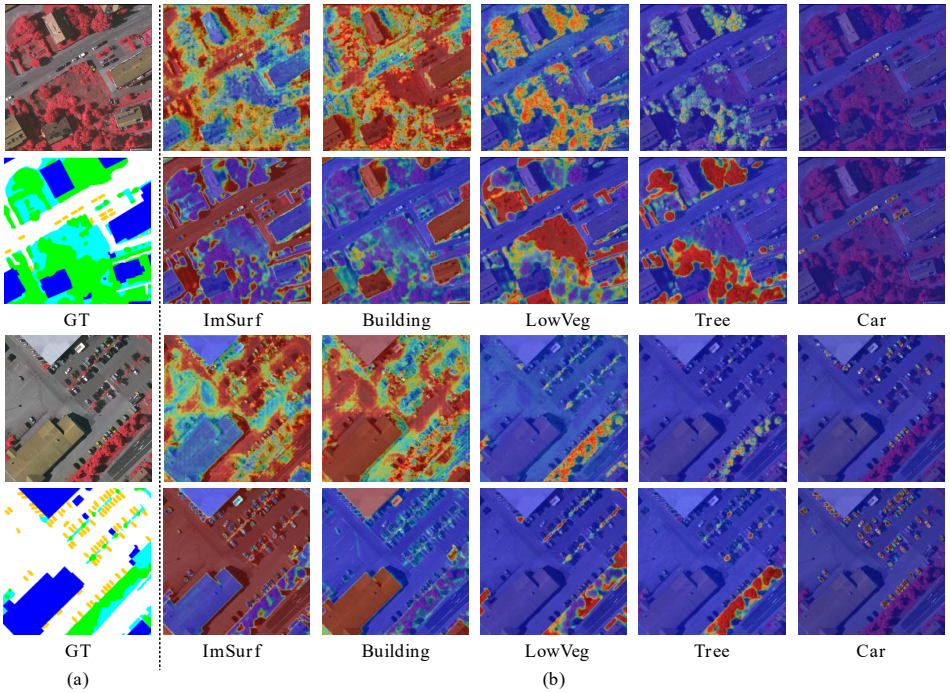

Figure 13. Visualization heatmaps of different category features before and after ACE processing. In (**a**), the top row shows the original input images, and the bottom row displays the corresponding color-coded labels. In (**b**), the top row presents the visualization results before applying ACE, while the bottom row shows the results after applying ACE. The red part indicates the segmentation result of the corresponding category. The deeper the red, the higher the response of the model to the image region when identifying that category.

Table 12. Comparison of complexity between the proposed CFENet and other mainstream methods.

Method	Params (M)	FLOPs (G)	mIoU (%)
PSPNet [41]	33.07	583.97	80.27/83.62
DeepLabV3+ [57]	40.42	186.85	69.93/72.84
U-Net [11]	28.29	546.71	78.67/77.07
MANet [59]	35.86	310.68	82.53/85.05
ABCNet [58]	13.43	63.04	84.04/85.66
UNetFormer [27]	11.68	46.97	83.35/86.10
DC-Swin [16]	45.33	184.25	83.95/86.59
RS^{3}Mamba [29]	49.66	253.20	84.01/86.30
MSGCNet [30]	17.51	75.72	84.62/86.23
CMTFNet [31]	30.07	130.10	84.26/85.17
BEMS-UNetFormer [61]	20.10	84.20	83.10/86.12
LOGCAN++ [18]	31.03	200.76	84.13/85.46
CFENet (ours)	24.68	161.08	**84.77/86.62**

As shown in Table 12, CFENet has a moderate number of parameters, but its computational complexity is relatively high. This increase in complexity mainly stems from the incorporation of the FF module, which introduces additional multi-scale feature interactions and frequency-aware operations, leading to higher computational costs. Nevertheless, this design choice enables CFENet to capture richer contextual and frequency information, thereby significantly enhancing feature representation capability. Overall, CFENet maintains a mid-level model complexity while demonstrating strong performance.

5. Discussion

The strong performance of CFENet on the ISPRS Vaihingen and Potsdam datasets can be mainly attributed to the collaborative design of the WPPM, FF, ACE, and MFR modules. This synergy significantly enhances the model's adaptability to practical remote sensing scenarios, enabling stable segmentation of high-resolution images with complex spatial structures and pronounced scale variations. However, the experimental results also reveal certain limitations. Although the WPPM effectively preserves spatial structural information during the downsampling stage, its ability to model long-range dependencies remains constrained by the inherently local receptive fields of convolution operations, leaving room for improvement in capturing large-scale continuous objects and global contextual structures. In addition, while the FF module successfully alleviates semantic inconsistency among encoder features and improves feature fusion quality, the additional computational overhead introduced by this module increases the overall model complexity, which may affect real-time performance and deployment feasibility in practical remote sensing applications. Furthermore, since the experiments are conducted on only two relatively similar datasets, further investigation on more diverse remote sensing datasets would be valuable to better assess the model's generalization potential.

In the future, we plan to explore more efficient global context modeling strategies and further lightweight architectural designs, aiming to improve inference efficiency while maintaining high segmentation accuracy, as well as evaluating the model on a wider range of datasets to strengthen its practical applicability.

6. Conclusions

This paper presents a novel semantic segmentation method, termed CFENet, designed to address the limitations of insufficient multi-scale feature representation and inadequate contextual modeling in remote sensing scenes. First, the proposed WPPM can effectively enhance multi-scale feature extraction while preserving critical information and improving

computational efficiency. Second, the FF module is introduced to improve the semantic consistency of encoder features, thereby enhancing the effectiveness of multi-level feature fusion. Finally, the decoder integrates the ACE and MFR modules in a collaborative manner, further strengthening contextual modeling and multi-scale feature reconstruction, enabling accurate and efficient semantic segmentation in complex remote sensing scenarios. Experimental results on the ISPRS Vaihingen and Potsdam datasets show that CFENet achieves stable and satisfactory performance when evaluated against several representative mainstream methods, indicating the effectiveness of the proposed network architecture.

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

Funding: This research were funded by the National Natural Science Foundation of China grant number 12402344 and the Research Project of Hubei Provincial Department of Education grant number B2023064.

Data Availability Statement: The original contributions presented in this study are included in the article. The source code and trained models for CFENet will be made publicly available upon acceptance.

Acknowledgments: We would like to express our special thanks to Shufen Ruan, Quan Wan, Ruijuan Chen, Mengyang Hu, and Kunfang Song for their valuable insights and technical support throughout the research process. We also appreciate the feedback from the editors and reviewers of *Remote Sensing*, which greatly contributed to improving the quality of this paper. Finally, we thank all individuals who participated in this research.

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

References

1. Usamentiaga, R.; Lema, D.G.; Pedrayes, O.D.; Garcia, D.F. Automated surface defect detection in metals: A comparative review of object detection and semantic segmentation using deep learning. *IEEE Trans. Ind. Applicat.* **2022**, *58*, 4203–4213. [CrossRef]
2. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2636–2645.
3. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
4. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
5. Akkus, Z.; Galimzianova, A.; Hoogi, A.; Rubin, D.L.; Erickson, B.J. Deep learning for brain MRI segmentation: State of the art and future directions. *J. Digit. Imaging* **2017**, *30*, 449–459. [CrossRef] [PubMed]
6. Lan, M.; Rong, F.; Jiao, H.; Gao, Z.; Zhang, L. Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5626513. [CrossRef]
7. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
8. Guo, Y.; Jia, X.; Paull, D. Effective sequential classifier training for SVM-based multitemporal remote sensing image classification. *IEEE Trans. Image Process.* **2018**, *27*, 3036–3048. [CrossRef]
9. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 109–117.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015, pp. 234–241.

12. Mnih, V.; Hinton, G.E. Learning to label aerial images from noisy data. In Proceedings of the International Conference on Machine Learning, Edinburgh, Scotland, 26 June–1 July 2012; pp. 567–574.
13. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**, arXiv:1606.02585. [CrossRef]
14. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]
15. Alhichri, H.; Alajlan, N.; Bazi, Y.; Rabczuk, T. Multi-scale convolutional neural network for remote sensing scene classification. In Proceedings of the IEEE International Conference on Electro/Information Technology, Rochester, MI, USA, 3–5 May 2018; pp. 1–5.
16. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [CrossRef]
17. Ma, X.; Zhang, X.; Pun, M.O.; Liu, M. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403215. [CrossRef]
18. Ma, X.; Lian, R.; Wu, Z.; Guo, H.; Yang, F.; Ma, M.; Wu, S.; Du, Z.; Zhang, W.; Song, S. Logcan++: Adaptive local-global class-aware network for semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4404216. [CrossRef]
19. Li, W.; Liao, M.; Hua, G.; Zhang, Y.; Zou, W. Contextual Guidance Network for Real-Time Semantic Segmentation of Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 16188–16203. [CrossRef]
20. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180. [CrossRef]
21. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
22. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 5229–5238.
23. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
24. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
25. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502. [CrossRef]
26. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9167–9176.
27. 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 J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
28. Ding, H.; Xia, B.; Liu, W.; Zhang, Z.; Zhang, J.; Wang, X.; Xu, S. A novel mamba architecture with a semantic transformer for efficient real-time remote sensing semantic segmentation. *Remote Sens.* **2024**, *16*, 2620. [CrossRef]
29. Ma, X.; Zhang, X.; Pun, M.O. RS^{3}Mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6011405. [CrossRef]
30. Zeng, Q.; Zhou, J.; Tao, J.; Chen, L.; Niu, X.; Zhang, Y. Multiscale global context network for semantic segmentation of high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622913. [CrossRef]
31. 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. [CrossRef]
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
34. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
35. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 603–612.
36. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
37. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

38. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
39. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
40. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
41. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
42. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
43. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19529–19539.
44. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819. [CrossRef]
45. Dong, Z.; Liu, T.; Gu, Y. Spatial and semantic consistency contrastive learning for self-supervised semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5621112. [CrossRef]
46. Zhang, J.; Liu, R.; Shi, H.; Yang, K.; Reiß, S.; Peng, K.; Fu, H.; Wang, K.; Stiefelhagen, R. Delivering arbitrary-modal semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023;; pp. 1136–1147.
47. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
48. Li, Y.; Li, Z.; Liu, H.; Wang, Q. ZMNet: Feature fusion and semantic boundary supervision for real-time semantic segmentation. *Vis. Comput.* **2025**, *41*, 1543–1554. [CrossRef]
49. Weng, X.; Yan, Y.; Chen, S.; Xue, J.H.; Wang, H. Stage-aware feature alignment network for real-time semantic segmentation of street scenes. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4444–4459. [CrossRef]
50. Chen, L.; Fu, Y.; Gu, L.; Yan, C.; Harada, T.; Huang, G. Frequency-aware feature fusion for dense image prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10763–10780. [CrossRef]
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Ni, Z.; Chen, X.; Zhai, Y.; Tang, Y.; Wang, Y. Context-guided spatial feature reconstruction for efficient semantic segmentation. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 239–255.
53. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
54. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the stars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5694–5703.
55. Ruan, J.; Xiang, S.; Xie, M.; Liu, T.; Fu, Y. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, Las Vegas, NV, USA, 6–8 December 2022; pp. 1150–1156.
56. 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 IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
57. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
58. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
59. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [CrossRef]
60. Li, X.; Tian, X.; Wang, Z.; Zhang, F.; Zhang, Y.; Yang, N.; Tian, C. SAM2Former: Segment Anything Model 2 Assisting UNet-Like Transformer for Remote Sensing Image Semantic Segmentation. *IEEE Access* **2025**, *13*, 115018–115032. [CrossRef]

61. Wang, J.; Chen, T.; Zheng, L.; Tie, J.; Zhang, Y.; Chen, P.; Luo, Z.; Song, Q. A multi-scale remote sensing semantic segmentation model with boundary enhancement based on UNetFormer. *Sci. Rep.* **2025**, *15*, 14737. [CrossRef]
62. Ma, X.; Wang, Z.; Hu, Y.; Zhang, X.; Pun, M.O. Kolmogorov-Arnold Network for Remote Sensing Image Semantic Segmentation. *arXiv* **2025**, arXiv:2501.07390. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.