remote sensing

MDPI

Article

WASP-Mamba: A Wavelet-Enhanced Mamba Framework for Remote Sensing Semantic Segmentation

Yuhao Zhang [1], Guolong Zhang [1], Yi Li [2], Yongbo Wu [1] and Zhiguo Zhou [1,*]

[1] School of Information Science and Technology, Northeast Normal University, Changchun 130117, China; zhangyh313@nenu.edu.cn (Y.Z.) zhanggl483@nenu.edu.cn (G.Z.) wuyb520@nenu.edu.cn (Y.W.)
[2] School of Computer Science, Changchun Humanities and Sciences College, Changchun 130117, China; liyi530@nenu.edu.cn
* Correspondence: zhouzg281@nenu.edu.cn

Highlights

What are the main findings?

- This paper proposes a novel WASP-Mamba semantic segmentation architecture that integrates multi-scale global modeling with a wavelet-inspired detail recovery mechanism to enhance feature representation in complex remote sensing scenes.
- Experiments on the LoveDA, ISPRS Vaihingen and ISPRS Potsdam datasets show that WASP-Mamba consistently outperforms state-of-the-art methods in multi-class land cover recognition, small object detection, and boundary preservation.

What are the implications of the main findings?

- The proposed multi-scale state-space modeling and adaptive frequency-domain decoupling effectively address scale variation while balancing global context modeling and local detail recovery, demonstrating the effectiveness of introducing frequency-aware mechanisms in upsampling.
- WASP-Mamba provides a robust solution for high-resolution remote sensing image segmentation, with strong potential for Earth observation applications.

Abstract

Semantic segmentation of high-resolution remote sensing (HRRS) imagery is a fundamental task in Earth observation but remains challenging due to severe scale variations and complex boundary structures. Although State Space Models (SSMs), particularly Mamba, have shown strong potential for efficient global modeling with linear complexity, existing methods are still limited in capturing multi-scale context and fine-grained spatial details. To address these issues, we propose a novel framework named WASP-Mamba for HRRS image segmentation. A Mamba-based spatial pyramid pooling module is introduced at the encoder bottleneck to enhance multi-scale feature aggregation and improve robustness to scale variations. In addition, a wavelet-inspired Mamba decoder is designed to decouple and reconstruct low-frequency semantic information and high-frequency details, mitigating boundary degradation caused by conventional upsampling. Extensive experiments on the LoveDA, ISPRS Vaihingen and ISPRS Potsdam datasets demonstrate that the proposed method achieves 53.90%, 84.32% and 87.06% mIoU, respectively. Compared with recent state-of-the-art methods, WASP-Mamba achieves superior segmentation performance.

Keywords: state space models; Mamba; high-resolution remote sensing; multi-scale feature; boundary preservation

Check for updates

Academic Editors: Xudong Kang, Sheng Xu, Qiaolin Ye and Yu Shen

Received: 19 April 2026
Revised: 22 June 2026
Accepted: 2 July 2026
Published: 4 July 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.

1. Introduction

High-resolution remote sensing (HRRS) imagery provides rich spatial and semantic information and has become an essential data source for a wide range of geospatial applications such as urban planning [1] and environmental monitoring [2]. Semantic segmentation of HRRS images aims to assign a semantic label to each pixel, enabling fine-grained scene understanding for downstream Earth observation tasks. However, unlike natural images, HRRS data exhibit highly complex spatial characteristics, including significant intra-class scale variations, high inter-class spectral similarity, dense object distributions, and weak or ambiguous boundary structures. These properties make pixel-level classification particularly challenging and limit the robustness of existing segmentation methods.

Recent advances in deep learning have significantly promoted the development of remote sensing image segmentation [3–6]. Convolutional neural network [7] (CNN)-based methods, such as U-Net [8], typically adopt encoder–decoder architectures with skip connections to capture hierarchical local features. However, CNNs are fundamentally constrained by local receptive fields, which limits their ability to model long-range dependencies and results in insufficient global contextual understanding, especially in large-scale and spatially heterogeneous remote sensing scenes. To address this limitation, Transformer-based [9] methods introduce self-attention mechanisms to model global relationships and achieve strong representation capability. Nevertheless, the quadratic computational complexity of self-attention leads to substantial memory and computational overhead when processing high-resolution remote sensing images, where dense pixel-wise interactions are required. This limitation restricts their scalability in practical large-scale Earth observation applications.

Recently, State Space Models (SSMs) [10], particularly Mamba, have emerged as an efficient alternative due to their linear computational complexity and strong long-range dependency modeling capability [11,12]. Several Mamba-based frameworks have demonstrated promising performance in visual recognition and remote sensing tasks. However, existing Mamba-based segmentation architectures still face two key limitations when applied to HRRS imagery. First, most methods lack explicit multi-scale context aggregation mechanisms at the bottleneck stage, making it difficult to effectively model objects with large-scale variations. Second, the decoding process in many existing designs still relies on conventional spatial-domain upsampling strategies, which tend to suppress high-frequency details and lead to small-object omission.

Motivated by these observations, we hypothesize that effective HRRS semantic segmentation requires not only efficient long-range dependency modeling but also explicit multi-scale contextual representation and frequency-aware feature reconstruction. In particular, multi-scale contextual modeling is essential for handling objects with varying spatial extents and improving structural consistency, while preserving fine-grained information is crucial for recovering detailed object appearance.

To validate this hypothesis, we propose a novel Mamba-based semantic segmentation framework, termed WASP-Mamba. The proposed framework consists of two key components. First, a Mamba-based Spatial Pyramid Pooling (M-SPP) module is introduced at the bottleneck stage to enhance multi-scale context aggregation and long-range dependency modeling. Second, a Wavelet-inspired Mamba Decoder (WMD) is designed to perform adaptive frequency-aware feature reconstruction, improving the recovery of fine-grained structures during decoding and alleviating information loss caused by conventional upsampling operations.

The main contributions of this work are summarized as follows:

(1) We propose a novel Mamba-based semantic segmentation framework, termed WASP-Mamba, which jointly enhances multi-scale contextual modeling and detail-preserving feature reconstruction for high-resolution remote sensing images.
(2) We design a Mamba-based Spatial Pyramid Pooling (M-SPP) module that integrates multi-scale dilated convolutions with state-space modeling to enhance contextual representation across different receptive fields and improve the modeling of complex spatial structures.
(3) We propose a Wavelet-inspired Mamba Decoder (WMD), which introduces adaptive frequency-aware reconstruction to preserve structural details and alleviate information loss during feature reconstruction, while maintaining boundary quality during decoding.
(4) Extensive experiments conducted on the LoveDA [13], ISPRS Vaihingen and ISPRS Potsdam datasets demonstrate the effectiveness and superiority of the proposed framework over several state-of-the-art methods.

2. Related Works

2.1. Methods Based on CNNs and Transformers

Convolutional neural networks (CNNs) [7] have been widely used in remote sensing semantic segmentation due to their strong local feature extraction capability. FCN [14] introduced dense pixel-wise prediction, while U-Net [8] and its variants such as ResUNet [15] enhanced feature fusion via encoder–decoder structures and residual learning. To address scale variations in high-resolution remote sensing images, DeepLab [16] series methods employed atrous convolutions and spatial pyramid pooling for multi-scale context modeling. Beyond conventional CNN-based segmentation models, pyramid structures have also been successfully integrated into advanced generative architectures to capture multi-scale spatial representations in remote sensing imagery [17]. However, CNN-based methods are inherently limited by their local receptive fields, restricting global context modeling.

To overcome this limitation, Transformer-based [9] methods introduce self-attention for long-range dependency modeling. Vision Transformer (ViT) [18] demonstrates the effectiveness of pure Transformer architectures, while Swin Transformer [19] improves efficiency through hierarchical windows. Hybrid models such as TransUNet [20] and UNetFormer [21] combine CNNs and Transformers to integrate local and global representations. Nevertheless, Transformer-based methods suffer from high computational cost and may still lose fine-grained spatial details in high-resolution remote sensing images.

2.2. Methods Based on Mamba

Recently, state space models (SSMs) [10], particularly Mamba, have emerged as a promising alternative for modeling long-range dependencies with theoretically linear complexity with respect to input sequence length. Unlike self-attention mechanisms, Mamba achieves global context modeling through selective state propagation, enabling efficient sequence modeling in high-resolution visual tasks.

In remote sensing image segmentation, several studies have explored Mamba-based architectures. RS^{3}Mamba [22] adopts a dual-branch design to balance local feature extraction and global dependency modeling. UNetMamba [23] integrates Mamba modules into a U-shaped encoder–decoder architecture, replacing traditional self-attention with efficient sequence modeling to enhance global representation capability. CM-UNet [24] further combines CNN-based local feature extraction with Mamba-based global modeling, improving robustness in complex spatial structures and large-scale variations.

Despite these advantages, Mamba-based methods still face challenges in effectively preserving fine-grained spatial details and boundary-sensitive information, particularly in

complex high-resolution remote sensing scenes. This limitation indicates that purely global sequence modeling is insufficient for accurate pixel-level segmentation.

2.3. Methods Based on Frequency-Domain

Frequency-domain analysis has recently attracted increasing attention in visual representation learning due to its ability to capture structural, textural, and edge-related information that is often less apparent in the spatial domain. Methods such as FTransformer [25] introduce frequency-aware attention mechanisms to enhance feature representation from a spectral perspective. Similarly, MIFNet [26] incorporates frequency-guided feature fusion to improve robustness in remote sensing image segmentation tasks.

Wavelet-based approaches further extend frequency analysis by decomposing images into multi-scale–frequency sub-bands. For example, WgANet [27] leverages wavelet transform to extract high-frequency components and enhances boundary representation through attention-based refinement, demonstrating the effectiveness of frequency-aware modeling for edge-sensitive segmentation.

Although these methods improve local structure and boundary preservation, they are typically used as auxiliary feature enhancement strategies and are rarely integrated with efficient global modeling frameworks such as state space models or Mamba. As a result, they are limited in jointly modeling long-range dependencies and fine-grained spatial details within a unified framework.

2.4. Summary

In summary, existing remote sensing semantic segmentation methods face several key limitations. CNN-based methods are effective for local feature extraction but lack global context modeling capability. Transformer-based methods improve global dependency modeling but suffer from high computational complexity and limited efficiency in high-resolution scenarios. Mamba-based methods provide an efficient alternative for long-range dependency modeling but still struggle to preserve fine-grained spatial details and boundary structures. Meanwhile, frequency-domain and wavelet-based methods enhance texture and edge representations but are not sufficiently integrated with modern efficient global modeling architectures.

To this end, we propose WASP-Mamba, a framework for high-resolution remote sensing image segmentation that integrates efficient global dependency modeling with frequency-aware feature enhancement to achieve accurate segmentation in complex scenes.

3. Materials and Methods

3.1. Overall Framework

WASP-Mamba employs a classic U-shaped encoder–decoder architecture, as shown in Figure 1. The encoder adopts ResT-Lite as the backbone network, following the setting in UNetMamba [23] for fair comparison. ResT-Lite [28] is a lightweight hierarchical Transformer designed for efficient multi-scale feature extraction in high-resolution remote sensing images. Compared with heavier backbones such as Swin Transformer and ConvNeXt, it offers a better balance between representation capability and computational efficiency. Therefore, it is used as a standard backbone in this work, allowing the focus to remain on the proposed M-SPP and WMD modules. At the encoder bottleneck, a Mamba Spatial Pyramid Pooling (M-SPP) module is introduced to capture multi-scale contextual information. In the decoder, the proposed WMD module performs feature decoupling and high-resolution reconstruction in the frequency domain. In addition, a Local Supervision Module (LSM) is adopted to facilitate the learning of intermediate representations through

multi-scale auxiliary supervision, as shown in Figure 2. This module follows the original design and is kept unchanged in this work.

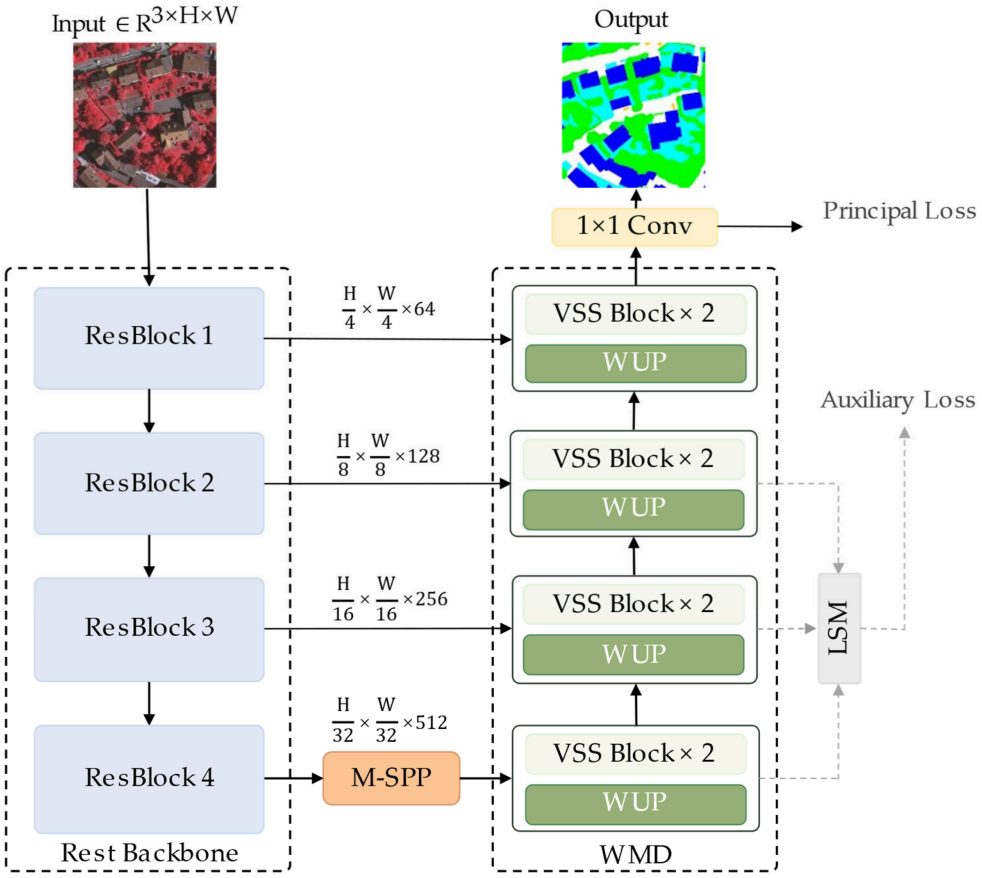


Figure 1. The architecture of the proposed WASP-Mamba framework.

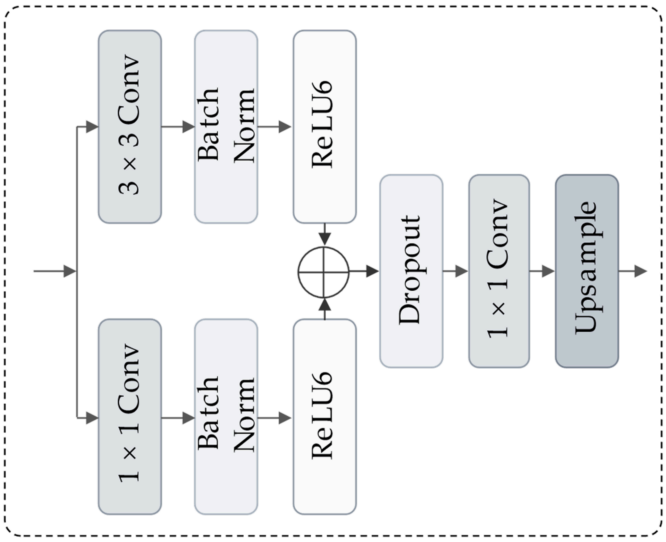


Figure 2. Local Supervision Module (LSM).

3.2. Mamba Spatial Pyramid Pooling (M-SPP)

In HRRS imagery, ground objects often exhibit significant scale variations, such as the coexistence of small vehicles and large agricultural fields, posing challenges to multi-scale perception. However, existing methods (e.g., UNetMamba [23]) typically rely on single-scale features for information propagation at the bottleneck stage of the encoder. Due to

the lack of multi-scale feature modeling and fusion, these methods may fail to preserve the internal structures of large objects or miss small-scale objects.

Inspired by the classical Atrous Spatial Pyramid Pooling (ASPP) module in the DeepLab [16] series, this paper draws upon its design concept of capturing information from varying receptive fields through multi-scale dilated convolutions to enhance multi-scale spatial feature representation. Based on this multi-scale feature fusion strategy, the VSS module is further introduced based on the state space model (Mamba), resulting in the proposed M-SPP module, as shown in Figure 3. The M-SPP module is placed at the deepest bottleneck stage of the encoder to better capture high-level contextual information. The VSS module is shown in Figure 4.

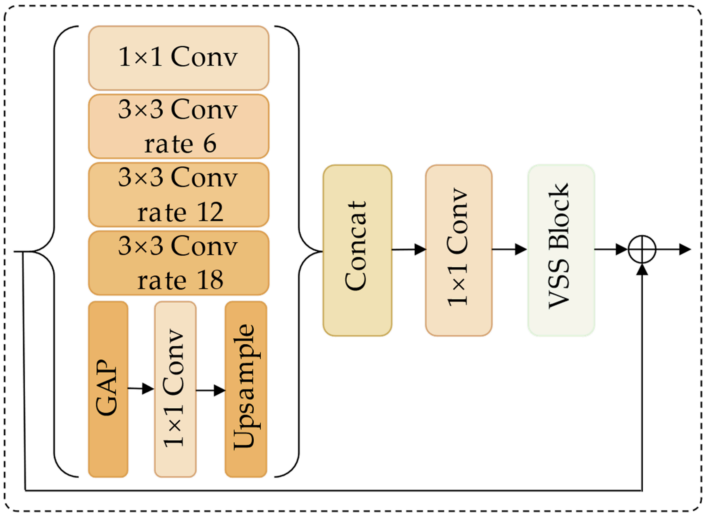


Figure 3. Mamba Spatial Pyramid Pooling (M-SPP) module.

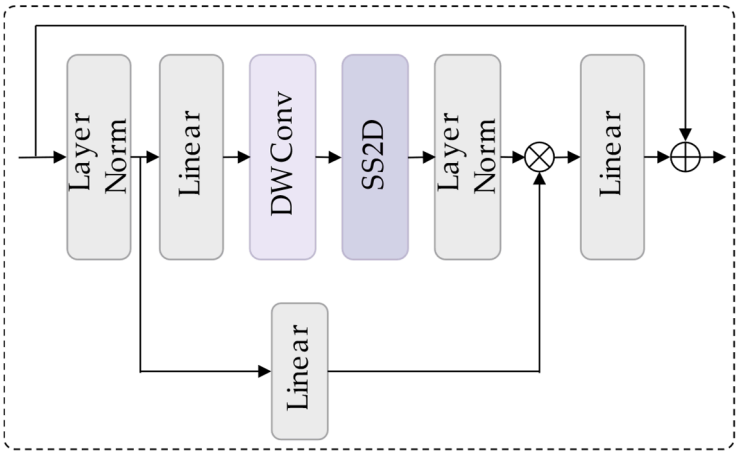


Figure 4. Visual State Space (VSS) module.

The VSS Block serves as the basic feature extraction unit in our framework. It integrates local spatial modeling and global context aggregation through depthwise convolution and the SS2D module, respectively. Meanwhile, a gating mechanism and residual connection are employed to enhance feature representation and optimization stability. This design enables effective modeling of both local details and long-range dependencies.

Let the bottleneck feature map output from the deepest layer of the encoder be denoted as $X_b \in R^{C_b \times H_b \times W_b}$, where C_b, H_b, and W_b denote the number of channels, height, and width, respectively. Due to the hierarchical downsampling operations in the encoder, H_b and W_b are smaller than the spatial size of the input image. The M-SPP module first

performs multi-scale modeling on X_b through five parallel branches to capture contextual information across different receptive field ranges.

Specifically, the first branch extracts global contextual information through Global Average Pooling (GAP), followed by a 1×1 convolution and bilinear upsampling:

$$X_{global} = \text{Upsample}\left(\text{Conv}_{1\times 1}(\text{GAP}(X_b))\right) \quad (1)$$

where $\text{GAP}(\cdot)$ compresses the spatial dimensions from $H_b \times W_b$ to 1×1 and $\text{Conv}_{1\times 1}(\cdot)$ performs feature projection. The resulting feature map is then restored to the original spatial resolution through bilinear upsampling, yielding the global contextual feature $X_{global} \in R^{C_b \times H_b \times W_b}$.

The second branch uses a 1×1 convolution to enhance local representation:

$$X_{local} = \text{Conv}_{1\times 1}(X_b) \quad (2)$$

where X_{local} denotes the output feature of the local branch, and $X_{local} \in R^{C_b \times H_b \times W_b}$.

The remaining three branches consist of 3×3 dilated convolutions with dilation rates of $r_i \in \{6, 12, 18\}$. These branches expand the receptive field without changing the spatial resolution, thereby capturing multi-scale contextual information from different receptive field ranges:

$$X_{r_i} = \text{DConv}_{3\times 3}^{r_i}(X_b), r_i \in \{6, 12, 18\} \quad (3)$$

where $\text{DConv}_{3\times 3}^{r_i}(\cdot)$ denotes a 3×3 dilated convolution with dilation rate r_i, and X_{r_i} denotes the output feature of the corresponding dilated convolution branch. Since padding is used to maintain the spatial resolution and the channel number remains unchanged in each branch, $X_{r_i} \in R^{C_b \times H_b \times W_b}$.

Subsequently, the features from the five branches are concatenated along the channel dimension, resulting in a feature map with $5C_b$ channels. A 1×1 convolution is then applied to fuse the cross-scale features and restore the channel number to C_b. Finally, the fused multi-scale feature X_{ms} is obtained:

$$X_{cat} = \text{Concat}\left(X_{global}, X_{local}, X_6, X_{12}, X_{18}\right), X_{cat} \in R^{5C_b \times H_b \times W_b} \quad (4)$$

$$X_{ms} = \text{Conv}_{1\times 1}(X_{cat}), X_{ms} \in R^{C_b \times H_b \times W_b} \quad (5)$$

where X_{cat} denotes the concatenated multi-branch feature, and X_{ms} denotes the fused multi-scale feature after channel fusion.

The fused multi-scale feature X_{ms} is mainly derived from local operations, which limits its ability to capture long-range spatial dependencies. To address this issue, X_{ms} is further fed into the VSS Block, which employs a 2D selective scan (SS2D) state-space modeling mechanism to transform the feature map into multi-directional sequences. Based on the state-space formulation, long-range information propagation can be effectively achieved, enabling the modeling of spatial relationships across different regions. This helps establish contextual associations between targets with significant scale variations (e.g., a small vehicle and a distant large building). Finally, a residual connection is introduced to fuse the original multi-scale features with the globally modeled features. The overall computation can be expressed as follows:

$$X_{msp} = \text{VSSM}(X_{ms}) + X_{ms} \quad (6)$$

where $X_{msp} \in R^{C_b \times H_b \times W_b}$ denotes the output feature of the M-SPP module.

3.3. Wavelet-Inspired Mamba Decoder (WMD)

In the feature decoding stage, most UNet-based architectures rely on deconvolution or bilinear interpolation for feature upsampling, which may lead to the loss of fine details such as building boundaries. To address this issue, this paper proposes a Wavelet-inspired Mamba Decoder (WMD). In each decoding layer, the WMD consists of two cascaded components: a Wavelet-inspired Upsampling module (WUP) (As shown in Figure 5) for detail-preserving resolution restoration and a VSS Block for spatial context refinement.

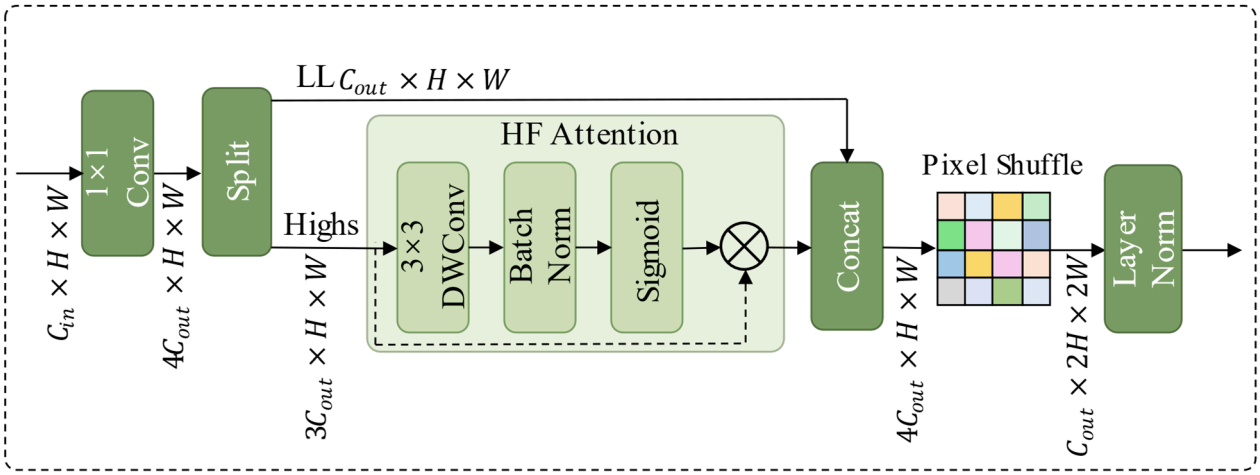


Figure 5. Wavelet-inspired Upsampling Module.

From the perspective of frequency-domain reconstruction, resolution restoration can be achieved by the two-dimensional inverse discrete wavelet transform (IDWT) [29,30]. However, IDWT usually relies on predefined fixed filters for signal reconstruction, which limits its adaptability to diverse and complex remote sensing scenes. To address this limitation, we propose WUP, which learns a data-driven wavelet-like decomposition and reconstruction process. Unlike traditional discrete wavelet transform methods, WUP does not rely on predefined wavelet bases or fixed filters, but instead learns the decomposition in a fully data-driven manner. Given the low-resolution decoder feature $X_d \in R^{C_{in} \times H \times W}$ from the previous layer, where C_{in}, H, and W denote its channel number, height, and width, respectively, WUP first generates low-frequency-like and high-frequency-like feature representations through a learnable 1×1 convolutional projection. Then, the spatial resolution is restored by Pixel Shuffle rearrangement. This design enables WUP to effectively preserve structural information and alleviate detail loss during feature reconstruction.

Specifically, given the input feature $X_d \in R^{C_{in} \times H \times W}$, WUP first applies a 1×1 convolution to project it into a $4C_{out}$ channel feature space, where $C_{out} = C_{in}/2$:

$$X_p = \mathrm{Conv}_{1 \times 1}(X_d),\ X_p \in R^{4C_{out} \times H \times W} \tag{7}$$

Then, the projected feature X_p is split along the channel dimension into a low-frequency-like component $X_{LL} \in R^{C_{out} \times H \times W}$ and high-frequency-like components $X_{Highs} \in R^{3C_{out} \times H \times W}$:

Here, X_{LL} is used to preserve coarse semantic structures, while X_{Highs} represents high-frequency details such as edges, textures, and object boundaries. To further enhance high-frequency information, a High-Frequency Attention Module (HFA) is introduced:

$$\hat{X}_{Highs} = X_{Highs} \otimes \sigma(\mathrm{BN}(\mathrm{DWCon}\, v_{3 \times 3}\, (X_{Highs}))) \tag{8}$$

where $\hat{X}_{Highs}$ denotes the enhanced high-frequency feature, $DWConv_{3 \times 3}(\cdot)$ denotes a 3 $\times$ 3 depthwise convolution, BN($\cdot$) denotes Batch Normalization, $\sigma(\cdot)$ denotes the sigmoid activation function, and $\otimes$ denotes element-wise multiplication.

The low-frequency-like component X_{LL} is concatenated with the enhanced high-frequency-like component $\hat{X}_{Highs}$, and the concatenated feature is then rearranged by Pixel Shuffle (PS) to obtain the upsampled feature:

$$X_{up} = PS(Concat(X_{LL}, \hat{X}_{Highs})),\ X_{up} \in R^{C_{out} \times 2H \times 2W} \quad (9)$$

where PS($\cdot$) denotes the Pixel Shuffle operation with an upsampling factor of 2. Specifically, the concatenated feature has a dimension of $4C_{out} \times H \times W$, and Pixel Shuffle rearranges it into $C_{out} \times 2H \times 2W$, thereby increasing the spatial resolution while reducing the channel dimension.

Following upsampling, X_{up} is concatenated with the skip connection feature $X_s \in R^{Cout \times 2H \times 2W}$ from the corresponding encoder layer. Then, a linear projection is applied to fuse the concatenated feature and restore the channel number to C_{out}:

$$X_f = Linear(Concat(X_{up}, X_s)),\ X_f \in R^{C_{out} \times 2H \times 2W} \quad (10)$$

where X_s denotes the skip-connection feature from the encoder, and X_f denotes the fused feature after skip connection fusion. This operation combines deep semantic information with shallow spatial details under spatial alignment.

Finally, the fused feature X_f is fed into the VSS Block for spatial context refinement:

$$X_{dec} = VSSM(X_f),\ X_{dec} \in R^{C_{out} \times 2H \times 2W} \quad (11)$$

This operation, based on spatial alignment, the encoder (M-SPP) provides multi-scale semantic context, while the decoder focuses on frequency reconstruction to preserve spatial structures and boundary details. This design enables a complementary interaction where semantic information is injected without interfering with frequency reconstruction, ensuring a stable balance between global context and local detail preservation. Leveraging the 2D cross-scanning mechanism (SS2D) of the Mamba architecture, the VSS Block models long-range dependencies with linear complexity at high-resolution scales, alleviating semantic conflicts and enabling global context aggregation. The output is then passed to the next stage.

Although the VSS block is adopted in both the M-SPP and WMD modules, its functional role varies with the network stage. In the M-SPP bottleneck, where feature maps have low spatial resolution and rich semantic information, the VSS block primarily serves as a global semantic modeler to capture long-range dependencies after multi-scale context aggregation. In contrast, the VSS blocks in the WMD operate on progressively higher-resolution features and focus on detail refinement and spatial consistency during feature reconstruction. Moreover, VSS blocks at different stages do not share parameters. Since their input resolutions, channel dimensions, and optimization objectives differ, each VSS block is independently learned to better adapt to its specific functional role.

Compared with traditional IDWT-based upsampling methods, the proposed design does not rely on fixed filters and instead learns the frequency decomposition process in a data-driven manner, allowing better adaptation to diverse geometric structures in complex remote sensing scenes.

3.4. Loss Function

To optimize the proposed WASP-Mamba framework, we employ a joint loss function consisting of Cross-Entropy (CE) loss and Dice loss. The principal loss L_p is a combination of a Dice loss L_{dice} and a Cross-Entropy loss L_{ce}, the formulations are as follows:

$$L_p = L_{dice} + L_{ce} \tag{12}$$

The detailed formulations of L_{dice} and L_{ce} are defined as follows:

$$L_{dice} = 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{y_k^{(n)}\hat{y}_k^{(n)}}{y_k^{(n)} + \hat{y}_k^{(n)}} \tag{13}$$

$$L_{ce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \log \hat{y}_k^{(n)} \tag{14}$$

where N and K denote the number of samples and the number of categories, respectively. $y^{(n)}$ and $\hat{y}^{(n)}$ represent the one-hot encoding of the true semantic labels and the corresponding softmax output of the network. Specifically, for the sample $n \in [1, \ldots, N]$, $\hat{y}_k^{(n)}$ indicates the predicted confidence of belonging to category k.

Furthermore, to effectively achieve the local supervisory role and improve intermediate feature representations, the classic cross-entropy loss is also utilized as the auxiliary loss L_{aux} deployed on the auxiliary head. Eventually, to balance the effects of L_p and L_{aux} and achieve optimal performance, the overall loss function L for training WASP-Mamba is formulated as a weighted equation:

$$L = L_p + \alpha L_{aux} \tag{15}$$

where the weight factor α set to 0.4 [21].

4. Experiment Settings

4.1. Dataset

4.1.1. LoveDA Dataset

The LoveDA dataset consists of 5987 HRRS images [13], each with a spatial resolution of 1024 × 1024 pixels, covering seven semantic classes: buildings, roads, water, wasteland, forest, agricultural land, and background. The dataset includes both rural and urban scenes collected from three cities in China: Nanjing, Changzhou, and Wuhan. Due to the presence of multi-scale objects, complex backgrounds, and imbalanced class distribution, the dataset poses significant challenges for semantic segmentation. In our experiments, 2522 images were used for training, 1669 for validation, and 1796 for testing, without additional cropping.

4.1.2. ISPRS Vaihingen Dataset

The ISPRS Vaihingen dataset consists of 33 high-resolution true orthophotos, with an average spatial resolution of 2494 × 2064 pixels, including three spectral bands: near-infrared (NIR), red (R), and green (G). The dataset covers six semantic classes: impermeable surfaces, buildings, low vegetation, trees, vehicles, and background. In our experiments, 17 images (IDs: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, and 33) were used for testing, while the remaining 16 images were used for training. All images were cropped into patches of size 1024 × 1024.

4.1.3. ISPRS Potsdam Dataset

The ISPRS Potsdam dataset consists of 38 true orthophoto (TOP) tiles with a fine spatial resolution, each covering an area of 6000 × 6000 pixels. It contains the same semantic categories as the Vaihingen dataset and provides four multispectral bands (red, green, blue, and near-infrared), along with digital surface model (DSM) and normalized digital surface model (nDSM) data. Following the standard experimental protocol, the image tiles with IDs 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, and 7_13 were used for testing, while the remaining 23 tiles were used for training. Tile 7_10 was excluded due to annotation errors. For a fair comparison with existing methods, only the RGB bands were utilized in our experiments. All images were cropped into patches of size 1024 × 1024.

4.2. Implementation Details

All experiments were implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU. The environment was configured with Python 3.8, CUDA 11.8, PyTorch 2.0.0, and Torchvision 0.15.1. The state-space modeling operations in WASP-Mamba relied on the mamba-ssm (v1.1.1) and causal-conv1d (v1.2.0) libraries with a custom-compiled selective-scan kernel. The AdamW optimizer was adopted with an initial learning rate of 0.0006 and a weight decay of 0.00025. For both datasets, models were trained for 100 epochs with a batch size of 8. During the training phase, all methods applied data augmentation techniques, including random scaling, horizontal and vertical flipping, and random rotation. During the testing phase, all methods employed a Test-Time Augmentation (TTA) strategy based on horizontal and vertical flipping.

4.3. Evaluation Metrics

To comprehensively evaluate the performance of the proposed model, three standard evaluation metrics are adopted, including overall accuracy (OA), mean Intersection over Union (mIoU), and F1 score.

The Intersection over Union (IoU) measures the overlap between the predicted and ground-truth regions and is defined as the ratio of the intersection to the union:

$$\mathrm{IoU_i} = \frac{\mathrm{TP_i}}{\mathrm{TP_i} + \mathrm{FP_i} + \mathrm{FN_i}} \tag{16}$$

where $\mathrm{TP_i}$, $\mathrm{FP_i}$, and $\mathrm{FN_i}$ denote the numbers of true positive, false positive, and false negative pixels for class i, respectively.

The mean Intersection over Union (mIoU) is defined as the average IoU across all classes:

$$\mathrm{mIoU} = \frac{1}{\mathrm{K}} \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{IoU_i} \tag{17}$$

where K denotes the total number of classes. The mIoU is a widely used evaluation metric in semantic segmentation, which effectively reflects the overall segmentation performance across different classes.

The F1 score is the harmonic mean of precision and recall and is used to evaluate the classification performance of the model. It is defined as follows:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{18}$$

Overall accuracy (OA) denotes the proportion of correctly classified pixels and is defined as:

$$\mathrm{OA} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{TP_i}}{\sum_{\mathrm{i}=1}^{\mathrm{K}} \left(\mathrm{TP_i} + \mathrm{FP_i}\right)} \tag{19}$$

5. Result

5.1. Comparative Experiments

To evaluate the effectiveness of the proposed method, a series of representative state-of-the-art semantic segmentation methods for remote sensing images are selected for comparison. These methods include CNN-based approaches (RECNet [31]), Transformer-based approaches (BANet [32], DC-Swin [33], and UNetFormer [21]), and state-space model (SSM)-based approaches (RS^{3}Mamba [22], UNetMamba [23], UMFormer [34], E-PyramidMamba [35], and PPMamba [36]).

5.1.1. Quantitative Analysis

The results on the LoveDA dataset are presented in Table 1, Bac., Bui., Roa., Wat., Bar., For., and Agr. denote Background, Building, Road, Water, Barren, Forest, and Agriculture, respectively. The proposed method achieves the best overall performance, reaching an mIoU of 53.90%. In terms of class-wise results, the proposed method achieves the best performance in the Background (48.39%), Forest (47.98%), and Agriculture (64.30%) categories. In the Road category, the proposed method achieves an IoU of 58.77%, outperforming all compared methods.

Table 1. Quantitative comparison results on the LoveDA dataset. All images are resized to 1024 × 1024 pixels. The best results are highlighted in bold.

Method	Backbone	Bac.	Bui.	Roa.	Wat.	Bar.	For.	Agr.	mIoU (%)
BANet [32]	ResT-Lite	46.40	57.71	53.74	80.20	15.75	45.36	63.72	51.84
DC-Swin [33]	Swin-T	44.36	58.65	56.41	80.72	18.18	45.79	59.38	51.93
UNetFormer [21]	ResNet-18	45.29	56.36	57.70	80.72	**21.27**	46.37	61.13	52.69
RS3Mamba [22]	R18-VMamba-T	45.64	59.13	57.84	79.45	15.11	45.75	63.36	52.32
UNetMamba [23]	ResT-Lite	46.63	**61.08**	56.15	81.18	17.89	46.79	62.62	53.19
UMFormer [34]	ResNet-18	44.87	57.40	55.44	80.45	14.84	45.39	63.01	51.63
E-PyramidMamba [35]	ResNet-18	44.93	57.81	55.13	80.74	17.46	47.03	62.89	52.28
PPMamba [36]	ResNet-18	45.86	60.10	57.08	**81.86**	19.41	46.34	63.61	53.47
RECNet [31]	ResNet-18	44.42	56.93	56.80	81.00	16.02	46.19	61.76	51.87
WASP-Mamba (Ours)	ResT-Lite	**48.39**	60.10	**58.77**	80.68	17.08	**47.98**	**64.30**	**53.90**

The results on the ISPRS Vaihingen dataset are presented in Table 2, where Imp., Bui., Low., Tre., and Car denote Impervious Surface, Building, Low Vegetation, Tree, and Car, respectively. WASP-Mamba achieves an mIoU of 84.32%, an mF1 of 91.31%, and an OA of 93.39%, outperforming all competing methods across the overall evaluation metrics. From a class-wise perspective, the proposed method obtains the highest IoU scores for the Tree (82.53%) and Car (79.51%) categories. In particular, the notable improvement in the Car category indicates the effectiveness of the proposed framework in capturing fine-grained details and segmenting small-scale objects.

The results on the ISPRS Potsdam dataset are presented in Table 3. WASP-Mamba achieves an mIoU of 87.06%, an mF1 of 92.95%, and an OA of 91.91%, outperforming all competing methods across the overall evaluation metrics. From a class-wise perspective, the proposed method obtains the highest IoU scores for the Impervious Surface (89.57%), Building (94.45%), and Low Vegetation (78.74%) categories. In particular, the notable improvements in the Building and Impervious Surface categories demonstrate the effectiveness of the proposed framework in aggregating multi-scale contextual information for large-scale urban features and preserving complex structural boundaries.

To address the limitation of region-based metrics, such as mIoU, which primarily evaluate pixel-wise classification accuracy and may not fully reflect boundary quality—

an important aspect in high-resolution remote sensing segmentation—we introduce the Boundary Intersection over Union (B-IoU) [37] metric as a complementary evaluation criterion to explicitly assess boundary delineation performance. For a fair comparison, B-IoU is computed for all competing methods under the same evaluation protocol. As reported in Table 4, WASP-Mamba achieves the highest mean B-IoU scores, reaching 51.21% on the ISPRS Vaihingen dataset and 49.64% on the ISPRS Potsdam dataset, consistently outperforming all compared methods.

Table 2. Quantitative comparison results on the ISPRS Vaihingen dataset. All images are resized to 1024 × 1024 pixels. The best results are highlighted in bold.

Method	Backbone	Imp.	Bui.	Low.	Tre.	Car.	mF1	mIoU (%)	OA
BANet [32]	ResT-Lite	93.31	90.75	72.72	82.04	77.95	90.73	83.35	93.01
DC-Swin [33]	Swin-T	93.84	91.72	73.39	82.48	75.51	90.72	83.39	93.32
UNetFormer [21]	ResNet-18	93.79	91.37	73.28	82.13	75.61	90.63	83.24	93.27
RS3Mamba [22]	R18-VMamba-T	93.77	91.69	72.94	82.11	77.53	90.86	83.61	93.27
UNetMamba [23]	ResT-Lite	93.62	91.47	73.11	81.99	78.66	90.98	83.77	93.19
UMFormer [34]	ResNet-18	93.37	90.41	72.91	81.53	75.75	90.38	82.80	92.92
E-PyramidMamba [35]	ResNet-18	93.59	91.20	72.87	81.73	74.75	90.38	82.83	93.08
PPMamba [36]	ResNet-18	93.74	91.66	73.46	82.42	75.47	90.70	83.35	93.29
RECNet [31]	ResNet-18	**94.13**	**92.08**	**74.22**	82.17	78.29	91.22	84.18	**93.52**
WASP-Mamba (Ours)	ResT-Lite	93.78	91.78	74.00	**82.53**	**79.51**	**91.31**	**84.32**	93.39

Table 3. Quantitative comparison results on the ISPRS Potsdam dataset. All images are resized to 1024 × 1024 pixels. The best results are highlighted in bold.

Method	Backbone	Imp.	Bui.	Low.	Tre.	Car.	mF1	mIoU (%)	OA
BANet [32]	ResT-Lite	87.33	92.51	77.17	79.72	90.46	92.03	85.44	90.86
DC-Swin [33]	Swin-T	89.37	93.60	78.17	**80.88**	92.43	92.86	86.89	91.68
UNetFormer [21]	ResNet-18	88.27	92.81	77.86	79.78	92.39	92.48	86.22	91.21
RS3Mamba [22]	R18-VMamba-T	87.71	93.14	77.81	80.12	92.75	92.53	86.32	91.21
UNetMamba [23]	ResT-Lite	88.51	93.44	77.88	80.04	91.40	92.50	86.25	91.34
UMFormer [34]	ResNet-18	88.80	93.78	78.31	79.48	93.14	92.74	86.70	91.51
E-PyramidMamba [35]	ResNet-18	88.41	93.09	77.72	80.42	**92.98**	92.65	86.53	91.36
PPMamba [36]	ResNet-18	88.48	92.99	77.80	**81.02**	92.96	92.73	86.65	91.40
RECNet [31]	ResNet-18	88.45	92.31	76.65	79.26	92.79	92.27	85.89	90.90
WASP-Mamba (Ours)	ResT-Lite	**89.57**	**94.45**	**78.74**	80.17	92.36	**92.95**	**87.06**	**91.91**

Table 4. Quantitative comparison results of Mean Boundary Intersection over Union on the ISPRS Vaihingen and ISPRS Potsdam datasets. All images are resized to 1024 × 1024 pixels. The best results are highlighted in bold.

Method	B-mIoU (%)	
	Vaihingen	Potsdam
BANet [32]	50.28	48.51
DC-Swin [33]	50.18	49.26
UNetFormer [21]	50.48	49.08
RS3Mamba [22]	50.64	48.96
UNetMamba [23]	51.09	49.44
UMFormer [34]	49.79	48.89
E-PyramidMamba [35]	50.00	48.87
PPMamba [36]	50.24	49.21
RECNet [31]	51.12	48.54
WASP-Mamba (Ours)	**51.21**	**49.64**

5.1.2. Qualitative Analysis

As shown in Figure 6, the qualitative segmentation results of WASP-Mamba are compared with those of several representative methods on the LoveDA dataset. Visual differences among the methods can be observed in the highlighted regions.

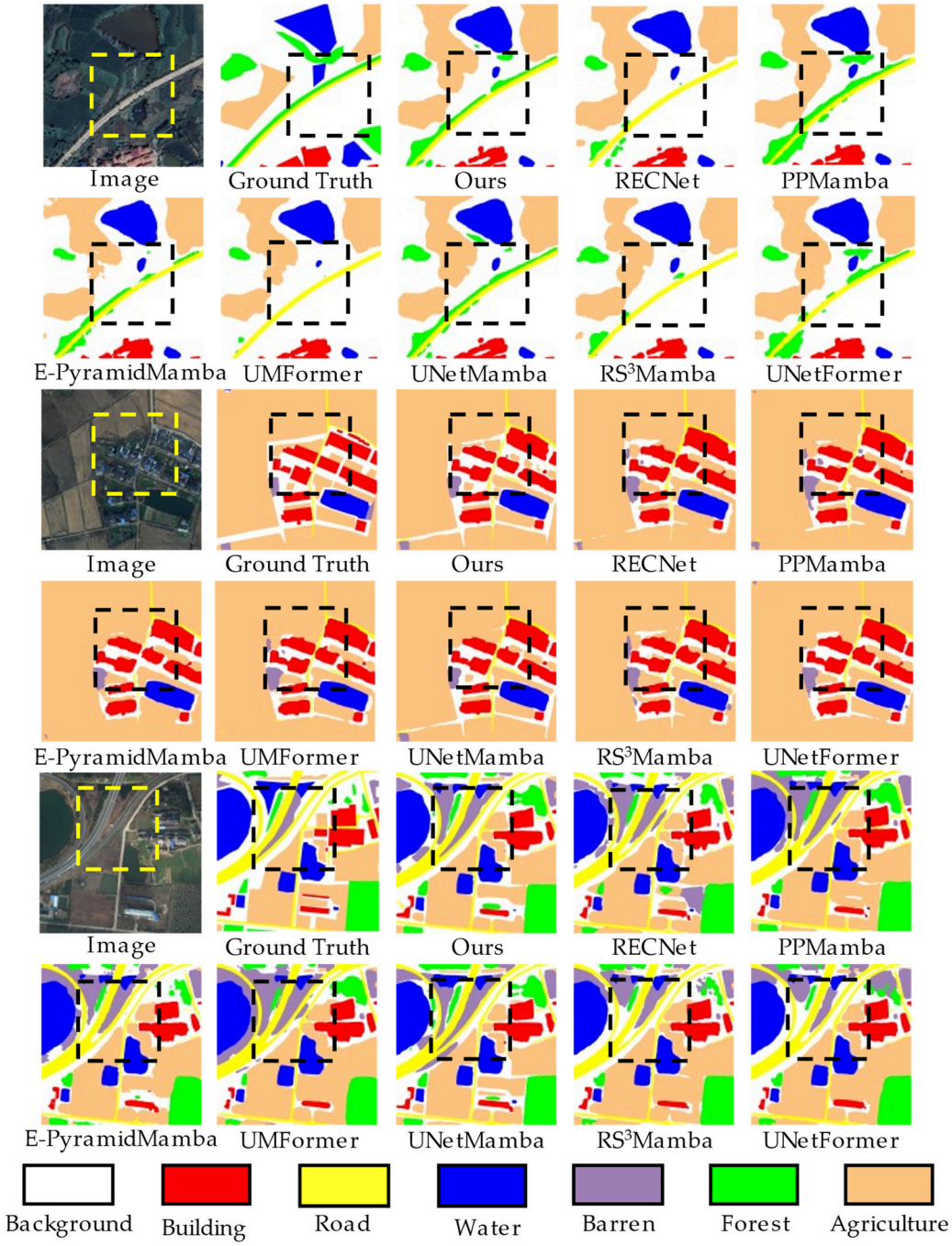


Figure 6. Qualitative comparison of different models on the LoveDA dataset. Black dashed boxes indicate areas of focus.

In scenes containing road–forest boundaries, the segmentation maps produced by different methods exhibit varying levels of continuity and boundary delineation. Compared with the other methods, WASP-Mamba generates road predictions that appear more continuous and exhibit clearer separation from adjacent vegetation regions. In agricultural land scenes, where large homogeneous regions dominate the image, noticeable differences can be observed in the completeness of the predicted agricultural areas. The segmentation

results of WASP-Mamba show a closer correspondence to the overall spatial distribution of the ground-truth annotations. In addition, in scenes containing large background regions, different methods exhibit varying prediction patterns in the highlighted areas. Compared with the competing approaches, WASP-Mamba produces segmentation maps with fewer visually observable category inconsistencies.

As shown in Figure 7, qualitative comparisons are further conducted on the ISPRS Vaihingen dataset. Visual inspection of the highlighted regions reveals observable differences among the compared methods.

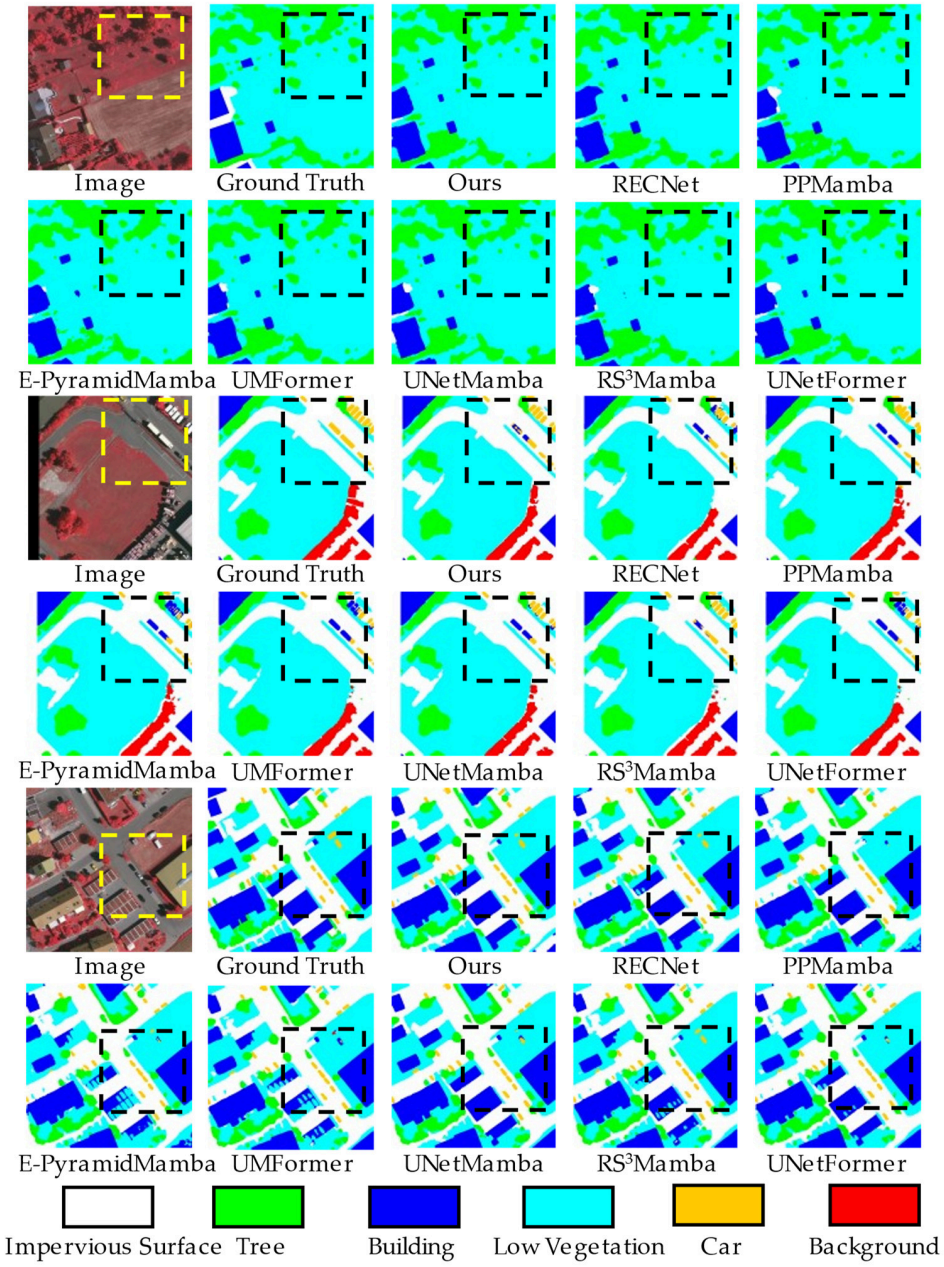


Figure 7. Qualitative comparison of different models on the ISPRS Vaihingen dataset. Black dashed boxes indicate areas of focus.

In scenes containing trees and low vegetation, the boundaries between the two categories are often complex and irregular. Compared with other methods, WASP-Mamba generates segmentation results that exhibit clearer delineation between these adjacent categories in the highlighted regions. For scenes containing impervious surfaces and vehicles, noticeable differences can be observed in the predicted vehicle regions. Compared with

several competing methods, WASP-Mamba identifies a larger number of vehicle instances, while the predicted object shapes remain visually consistent with the reference annotations. Furthermore, in building-dominated scenes, the segmentation results generated by different methods exhibit varying levels of agreement with the ground truth. In the highlighted regions, WASP-Mamba produces building predictions with clearer object boundaries and more complete structural representations.

As shown in Figure 8, qualitative comparisons are further conducted on the ISPRS Potsdam dataset. Visual inspection of the highlighted regions reveals observable differences among the compared methods.

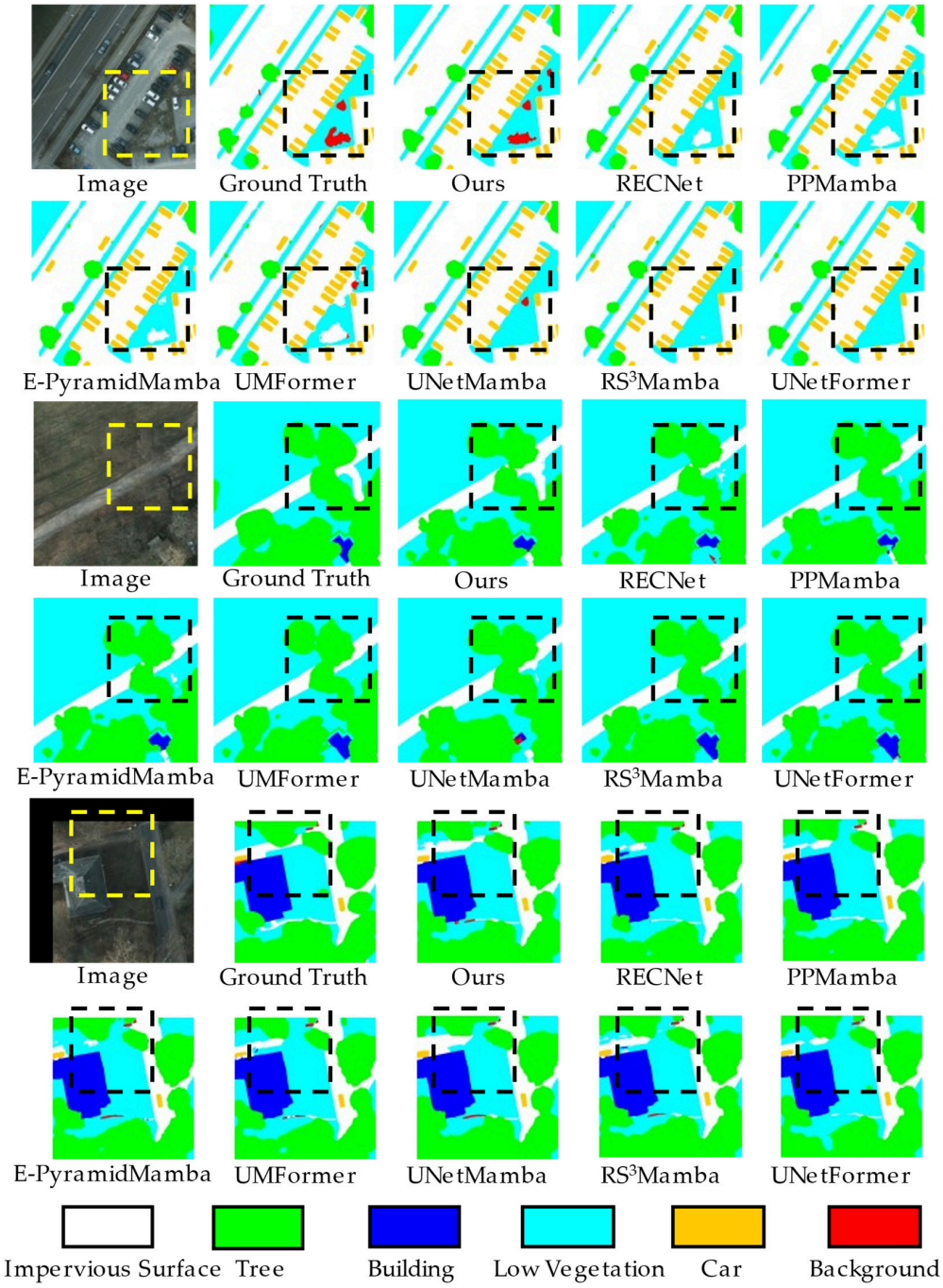


Figure 8. Qualitative comparison of different models on the ISPRS Potsdam dataset. Black dashed boxes indicate areas of focus.

In scenes containing complex backgrounds and low vegetation, distinguishing between these two categories is often challenging due to their potentially similar visual features. Compared with several competing methods that exhibit severe class confusion and irregular edges, WASP-Mamba produces segmentation results with a clear and accurate delineation between the background and low vegetation regions. Furthermore, in scenes featuring a complex mixture of impervious surfaces, low vegetation, and trees, the boundaries among these diverse land covers are highly prone to distortion during feature reconstruction. While other methods struggle to effectively separate the ground objects—often confusing the impervious surfaces with the surrounding greenery—WASP-Mamba effectively maintains the structural continuity of the impervious surfaces and precisely distinguishes them from the adjacent low vegetation. These visual improvements further demonstrate the superiority of the proposed framework in preserving high-frequency boundary details and handling highly easily confused categories.

5.2. Ablation Research

5.2.1. Ablation of the Model Structure

To verify the effectiveness of each component, ablation experiments were conducted on the LoveDA, ISPRS Vaihingen, and Potsdam datasets. Using the baseline model, we quantitatively evaluated the individual and combined contributions of the M-SPP and WMD modules. The results are summarized in Table 5.

Table 5. Ablation study results on the LoveDA, ISPRS Vaihingen, and ISPRS Potsdam datasets. Best values are shown in bold.

Method	mIoU (%)			B-mIoU (%)	
	LoveDA	Vaihingen	Potsdam	Vaihingen	Potsdam
Baseline	53.19	83.77	86.25	51.09	49.44
Baseline + M-SPP	53.34	84.15	86.46	**51.29**	**49.68**
Baseline + WMD	53.46	84.09	86.58	51.16	49.42
Baseline + M-SPP + WMD	**53.90**	**84.32**	**87.06**	51.21	49.64

When the M-SPP module is introduced, the mIoU increases from 53.19% to 53.34% on LoveDA, from 83.77% to 84.15% on ISPRS Vaihingen, and from 86.25% to 86.46% on Potsdam. Similarly, incorporating the WMD module improves the mIoU to 53.46%, 84.09%, and 86.58% on the three datasets, respectively.

When both M-SPP and WMD are integrated, the model achieves the highest performance on all three datasets, reaching 53.90% mIoU on LoveDA, 84.32% mIoU on ISPRS Vaihingen, and 87.06% mIoU on Potsdam. Compared with the baseline, the overall improvements are 0.71%, 0.55%, and 0.81%, respectively.

To further assess boundary quality, the Boundary IoU (B-mIoU) metric is also included in the ablation study. The M-SPP module improves the B-mIoU from 51.09% to 51.29% on Vaihingen and from 49.44% to 49.68% on Potsdam. When both M-SPP and WMD are incorporated, the model achieves B-mIoU scores of 51.21% on Vaihingen and 49.64% on Potsdam. Although these values are slightly lower than those obtained using M-SPP alone, the combined model consistently delivers the best overall segmentation performance in terms of mIoU.

These results indicate that both multi-scale contextual modeling and frequency-aware feature reconstruction contribute positively to segmentation performance. In particular, the improvements in B-mIoU suggest that effective contextual representation plays an important role in accurate boundary localization. Although the highest B-mIoU is achieved by the

M-SPP-only configuration, the integration of WMD further improves overall segmentation accuracy, resulting in the best overall performance across all evaluated datasets.

5.2.2. Ablation of the M-SPP Module

In order to further investigate the internal design of the M-SPP module, we conducted a specialized ablation study, and the results are shown in Table 6. Using only multi-scale dilated convolutions yields limited improvements and may introduce gridding artifacts due to sparse sampling. Incorporating the GAP branch further enhances performance by providing global contextual information, which complements the local multi-scale features extracted by the dilated convolution branches. Finally, the VSS module brings the most significant improvement by modeling long-range spatial dependencies and strengthening contextual interactions across different regions. The progressive performance gains observed in Table 6 confirm the effectiveness and complementarity of these three components in multi-scale feature representation and contextual modeling.

Table 6. Ablation study evaluating the internal design of the M-SPP module on the Vaihingen dataset; "√" indicates the presence of a component, while "×" indicates its absence. Best results are highlighted in bold.

Dilated Convs	GAP	VSS	Imp.	Bui.	Low.	Tre.	Car.	mIoU(%)
×	×	×	93.95	91.66	73.62	82.28	78.93	84.09
√	×	×	93.71	91.58	73.34	82.13	78.93	83.94
√	√	×	93.59	91.26	73.55	82.27	**80.03**	84.14
√	√	√	**93.78**	**91.78**	**74.00**	**82.53**	79.51	**84.32**

5.2.3. Ablation of the WMD Module

In order to further investigate the internal design of the WMD module, we conducted a dedicated ablation study to evaluate the specific contributions of the HFA components, and the results are shown in Table 7. The results show that removing the High-Frequency Attention (HFA) module leads to a noticeable performance drop, indicating that frequency decomposition alone may introduce redundant high-frequency responses that negatively affect feature reconstruction. By selectively enhancing informative high-frequency components while suppressing irrelevant noise, the HFA module facilitates the preservation of fine-grained structures and spatial details. The combination of frequency decomposition and HFA achieves the best performance, demonstrating their complementary roles in improving feature reconstruction and recovering detailed spatial information.

Table 7. Ablation study evaluating the internal design of the WMD module on the Vaihingen dataset; "√" indicates the presence of a component, while "×" indicates its absence. Best results are highlighted in bold.

Freq. Split & PS	HFA	Imp.	Bui.	Low.	Tre.	Car.	mIoU (%)
×	×	93.64	91.57	73.42	82.28	79.84	84.15
√	×	93.52	91.41	72.77	81.79	79.72	83.84
√	√	**93.78**	**91.78**	**74.00**	**82.53**	**79.51**	**84.32**

5.3. Complexity Analysis

To evaluate the computational efficiency of the proposed method, we compare Params, FLOPs, memory footprint, and inference speed (FPS), together with segmentation accuracy (mIoU), on the LoveDA dataset, as reported in Table 8.

Table 8. The computational complexity analysis is based on measurements obtained from 1024 × 1024 images. mIoU scores are reported on the LoveDA dataset. Bolded values indicate the best performance for each metric.

Method	Param (M)	FLOPs (G)	Memo (MB)	FPS (img/s)	mIoU (%)
BANet [32]	12.73	85.43	629.60	16.67	51.84
DC-Swin [33]	45.63	190.04	571.47	54.22	51.93
UNetFormer [21]	**11.72**	**47.03**	**307.23**	**256.33**	52.69
RS3Mamba [22]	43.32	157.89	1449.54	21.92	52.32
UNetMamba [23]	14.76	100.52	806.86	12.94	53.19
UMFormer [34]	12.37	48.55	474.04	75.61	51.63
E-PyramidMamba [35]	28.76	75.97	532.92	180.23	52.28
PPMamba [36]	21.70	84.64	969.88	15.42	53.47
RECNet [31]	17.96	125.11	653.48	72.45	51.87
WASP-Mamba (Ours)	27.85	114.32	1478.59	11.41	**53.90**

WASP-Mamba achieves the best segmentation performance with an mIoU of 53.90%, outperforming both CNN-based and Transformer-based methods, as well as recent Mamba-based variants. In terms of model complexity, WASP-Mamba contains 27.85M parameters and requires 114.32G FLOPs, compared with 14.76M parameters and 100.52G FLOPs for UNetMamba. This increase is mainly attributed to the additional learnable components introduced by the proposed M-SPP and WMD modules for multi-scale contextual modeling and frequency-aware feature reconstruction. Regarding inference efficiency, WASP-Mamba achieves 11.41 FPS, comparable to UNetMamba, while being slightly slower than highly optimized CNN/Transformer models, which is consistent with the computational characteristics of Mamba-based selective scanning.

Overall, WASP-Mamba demonstrates a favorable trade-off between segmentation accuracy and computational cost, highlighting the effectiveness of integrating multi-scale modeling with state space representations for high-resolution remote sensing segmentation.

5.4. Sensitivity Analysis of Dilation Rates in the M-SPP Module

To investigate the effect of dilation rates in the proposed M-SPP module, we conduct a sensitivity analysis on different dilation configurations, as shown in Table 9. Specifically, three commonly used settings are evaluated: (3, 6, 9), (6, 12, 18), and (9, 18, 27).

Table 9. Sensitivity analysis of different dilation rate settings in the M-SPP module. Bolded values indicate the best performance for each metric.

Dilation Rates	mIoU (%)	
	Vaihingen	LoveDA
(3, 6, 9)	**84.40**	53.54
(6, 12, 18)	84.32	**53.90**
(9, 18, 27)	84.19	53.02

The configuration (6, 12, 18) achieves the best overall performance on both the Vaihingen and LoveDA datasets, with mIoU scores of 84.32% and 53.90%, respectively. Although (3, 6, 9) yields slightly higher performance on the Vaihingen dataset, it performs worse on LoveDA. In contrast, larger dilation rates (9, 18, 27) lead to a noticeable performance drop, which may be attributed to overly sparse receptive fields.

Based on these empirical results, we adopt (6, 12, 18) as the default configuration for all experiments.

5.5. Sensitivity Analysis of the Loss Weight Coefficient in the Joint Loss Function

The weight coefficient α in the joint loss function is empirically set to 0.4, following commonly adopted settings in related studies. To further validate this choice and investigate its impact on optimization, we conduct a sensitivity analysis using different values of α, as reported in Table 10.

Table 10. Sensitivity analysis of the loss weight coefficient α. Bolded values indicate the best performance for each metric.

α	mIoU (%)	
	Vaihingen	**LoveDA**
0.2	84.19	53.17
0.4	**84.32**	**53.90**
0.6	84.09	53.33
0.8	83.74	53.29

The results indicate that $\alpha = 0.4$ achieves the best overall performance on both datasets. When α is assigned a smaller value (e.g., 0.2), the contribution of the auxiliary loss becomes relatively limited, resulting in degraded performance. In contrast, larger values of α (e.g., 0.6 and 0.8) also lead to performance deterioration, suggesting that excessive emphasis on the auxiliary objective may adversely affect the optimization of the main segmentation branch.

These observations indicate that an appropriate balance between the main loss and auxiliary loss is essential for effective training. In our experiments, $\alpha = 0.4$ provides the most favorable trade-off, yielding stable and consistent performance across both datasets. Therefore, $\alpha = 0.4$ is adopted in all subsequent experiments.

6. Discussion

The experimental results on the LoveDA, ISPRS Vaihingen and ISPRS Potsdam datasets demonstrate the capability of the proposed WASP-Mamba framework to handle challenging remote sensing scenes characterized by significant scale variations and complex spatial structures. Compared with existing encoder–decoder architectures, the proposed framework introduces the M-SPP and WMD modules to address limitations in multi-scale feature representation and feature reconstruction.

The improvements observed in the ablation studies suggest that the M-SPP module contributes to segmentation performance by introducing multi-scale contextual information at the bottleneck stage. Unlike conventional single-scale feature propagation, M-SPP aggregates features from multiple receptive fields before global state-space modeling, enabling richer contextual representation of objects with diverse scales and spatial layouts. The boundary ablation results further indicate that effective multi-scale contextual modeling can also contribute to more accurate boundary localization in complex remote sensing scenes.

The WMD module focuses on feature reconstruction during decoding. Conventional upsampling strategies will gradually weaken high-frequency information associated with object boundaries and small targets. By introducing a learnable high–low frequency decomposition strategy, WMD provides an alternative reconstruction mechanism that effectively helps preserve structural details during spatial resolution recovery. This observation is consistent with the qualitative results, where clearer object boundaries and more complete structures can be observed in several challenging regions.

Although M-SPP and WMD are connected through a standard encoder–decoder architecture, they operate at different stages of the segmentation pipeline. M-SPP primarily enhances contextual representation in deep features, whereas WMD focuses on detail

reconstruction during decoding. The ablation results show that the combination of these two modules achieves the best overall performance, suggesting that multi-scale contextual modeling and frequency-aware reconstruction provide complementary benefits.

Nevertheless, several limitations remain. The performance gains are not equally pronounced across all categories, and slight performance drops are observed for certain classes compared with some baselines. One possible explanation is that the design characteristics of M-SPP and WMD may affect different categories in different ways. For example, the frequency-aware reconstruction strategy in WMD emphasizes structural details and fine-scale information, which may introduce additional sensitivity to minor variations in relatively homogeneous regions such as water. Similarly, the multi-scale contextual aggregation performed by M-SPP may occasionally reduce the contribution of local geometric cues in categories characterized by regular and densely distributed structures, such as buildings. Furthermore, errors can still be observed in challenging scenarios, including confusion between trees and low vegetation, omission of extremely small vehicles, and boundary ambiguities in densely built urban regions. These observations suggest that distinguishing categories with highly similar spectral and spatial characteristics remains challenging for the proposed framework.

7. Conclusions

This paper addresses the issues of insufficient single-scale perception capability in semantic segmentation of HRRS images and the loss of details caused by upsampling in the decoding stage. A Mamba-based semantic segmentation model, WASP-Mamba, is proposed.

Specifically, an M-SPP module is introduced at the bottleneck layer to enhance multi-scale context representation by combining multi-scale dilated convolution with Mamba-based global modeling. In addition, a WMD module is designed in the decoding stage to perform adaptive frequency decomposition and reconstruction, thereby alleviating information loss and improving detail preservation during feature upsampling.

Experimental results on the LoveDA, ISPRS Vaihingen and ISPRS Potsdam datasets demonstrate that the proposed method consistently outperforms state-of-the-art approaches, achieving mIoU scores of 53.90%, 84.32% and 87.06%, respectively, which validates its effectiveness and robustness in complex remote sensing scenarios.

Future work will focus on developing more refined and efficient architectures, with particular emphasis on exploring scale–frequency interaction mechanisms, such as cross-scale and cross-frequency attention strategies or dynamic frequency-aware feature weighting, to better exploit the relationship between semantic scale information and frequency-domain representations.

Author Contributions: Conceptualization, Y.Z.; methodology, software, and formal analysis, Y.Z.; validation and investigation, G.Z. and Y.L.; resources, G.Z. and Y.L.; data curation and visualization, Y.W. and Y.L.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.L. and Y.W.; supervision, Z.Z.; project administration, Y.Z. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is publicly available. The LoveDA dataset is available at https://github.com/Junjue-Wang/LoveDA (accessed on 1 July 2026). The ISPRS Vaihingen and ISPRS Potsdam datasets are available at https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/default.aspx (accessed on 1 July 2026).

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

References

1. Lu, X.; Yuan, Y.; Zheng, X. Joint dictionary learning for multispectral change detection. *IEEE Trans. Cybern.* **2016**, *47*, 884–897. [CrossRef] [PubMed]
2. Gorry, B.; Sandino, J.; Moghadam, P.; Gonzalez, F.; Roberts, J. Advances in Machine Learning Approaches for UAV-Based Remote Sensing in Data-Deficient Antarctic Environments. *Remote Sens.* **2026**, *18*, 459. [CrossRef]
3. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
4. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
5. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
6. Ranjan, P.; Nandal, A.; Agarwal, S.; Kumar, R. A Dive into Generative Adversarial Networks in the World of Hyperspectral Imaging: A Survey of the State of the Art. *Remote Sens.* **2026**, *18*, 196. [CrossRef]
7. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **2002**, *86*, 2278–2324.
8. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
10. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
11. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063. [CrossRef]
12. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
13. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
14. 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.
15. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
16. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
17. Ranjan, P.; Ankur; Kumar, R. A Residual Pyramid-GAN Approach for Hyperspectral Image Classification. *Ann. Data Sci.* **2026**. [CrossRef]
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
19. 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/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
20. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
21. 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]
22. 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]
23. Zhu, E.; Chen, Z.; Wang, D.; Shi, H.; Liu, X.; Wang, L. Unetmamba: An efficient unet-like mamba for semantic segmentation of high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *22*, 6001205.
24. Liu, M.; Dan, J.; Lu, Z.; Yu, Y.; Li, Y.; Li, X. CM-UNet: Hybrid CNN-Mamba UNet for remote sensing image semantic segmentation. *arXiv* **2024**, arXiv:2405.10530.
25. Zhang, H.; Xie, G.; Li, L.; Xie, X.; Ren, J. Frequency-domain guided swin transformer and global-local feature integration for remote sensing images semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5612611. [CrossRef]
26. Fan, J.; Li, J.; Liu, Y.; Zhang, F. Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107638. [CrossRef]
27. Wang, Y.; Cao, L.; Deng, H. WgANet: A wavelet-guided attention network for remote sensing image semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *19*, 702–716.

28. Zhang, Q.; Yang, Y.-B. Rest: An efficient transformer for visual recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15475–15485.
29. Guo, T.; Seyed Mousavi, H.; Huu Vu, T.; Monga, V. Deep wavelet prediction for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 104–113.
30. Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; Zuo, W. Multi-level wavelet-CNN for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 773–782.
31. Li, Y.; Liu, G.; Bai, Y.; Zheng, C. Rotation Equivariant Convolution-Based Network for Semantic Segmentation of High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 6004705. [CrossRef]
32. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
33. 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]
34. 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. [CrossRef]
35. Wang, L.; Li, D.; Dong, S.; Meng, X.; Zhang, X.; Hong, D. PyramidMamba: Rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104884. [CrossRef]
36. Mu, J.; Zhou, S.; Sun, X. PPMamba: Enhancing semantic segmentation in remote sensing imagery by SS2D. *IEEE Geosci. Remote Sens. Lett.* **2024**, *22*, 6001705. [CrossRef]
37. Cheng, B.; Girshick, R.; Dollár, P.; Berg, A.C.; Kirillov, A. Boundary IoU: Improving object-centric image segmentation evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15334–15342.

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.