Original Paper

PPMamba: A Pyramid Pooling Local Auxiliary SSM-based Model for Remote Sensing Image Semantic Segmentation

Yin Hu, Xianping Ma, Jialu Sui and Man-On Pun[*]

School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China

ABSTRACT

Semantic segmentation is a vital task in the field of remote sensing (RS). However, conventional convolutional neural network (CNN) and transformer-based models face limitations in capturing long-range dependencies or are often computationally intensive. Recently, an advanced state space model (SSM), namely Mamba, was introduced, offering linear computational complexity while effectively establishing long-distance dependencies. Despite their advantages, Mamba-based methods encounter challenges in preserving local semantic information. To cope with these challenges, this paper proposes a novel network called Pyramid Pooling Mamba (PPMamba), which integrates CNN and Mamba for RS semantic segmentation tasks. The core structure of PPMamba, the Pyramid Pooling-State Space Model (PP-SSM) block, combines a local auxiliary mechanism with an omnidirectional state space model (OSS) that selectively scans feature maps from eight directions, capturing comprehensive feature information. Additionally, the auxiliary mechanism includes pyramid-shaped convolutional branches designed to extract features at multiple scales.

[*]Corresponding author: Man-On Pun, simonpun@cuhk.edu.cn.

Extensive experiments on three widely-used datasets, ISPRS Vai-
hingen, LoveDA Urban and WHU Buildings, demonstrate that
PPMamba achieves competitive performance compared to state-
of-the-art models. The source code will be made available at
https://github.com/KyotoSakura/PPMamba/.

Keywords: State space model, remote sensing, semantic segmentation

1 Introduction

The rapid development of remote sensing (RS) technologies has dramatically
transformed our understanding of time and space scales on the Earth. RS
technologies are extensively applied in agriculture [40], forestry [18], geology
[1] and environmental protection [42], enabling systematic analyses, assess-
ments, and predictions. Among these applications, semantic segmentation,
which assigns class labels to each pixel in an image, serves as a foundation
for many downstream geoscientific tasks, such as land cover classification and
urban expansion monitoring [7, 23].

In recent years, deep learning has significantly advanced the performance
of semantic segmentation in RS, primarily due to its ability to extract ab-
stract and hierarchically structured features from RS images [47]. Convolu-
tional neural network (CNN) and transformer are the most commonly used
techniques in state-of-the-art deep learning models. CNN-based models [13,
20, 41] excel at capturing local information through convolution operations,
while transformer-based models [27, 5, 8] leverage self-attention mechanisms
[35] to model long-distance dependencies. However, these methods still have
limitations in RS applications. CNN-based models struggle to capture global
context due to their restricted receptive fields, while transformers, although
capable of modeling long-range dependencies, face significant computational
challenges when handling high-resolution, large-scale RS data [8].

To overcome these challenges, Mamba, a novel state space model (SSM)-
based network, was introduced [11], offering a promising solution to effec-
tively capture long-distance dependencies with linear computational complex-
ity. Various SSM-based models have been successfully applied across differ-
ent domains, including Vmamba [21] and Vision Mamba [46] in computer
vision, as well as RSMamba [6] and RS3Mamba [25] in RS. Innovations such
as Mamba-in-Mamba [45] for hyperspectral image classification, Pan-Mamba
[15], and ChangeMamba [4] for RS pan-sharpening and change detection have
also emerged. Despite the advantages of these models, they struggle to char-
acterize local details, which is critical for accurate RS image segmentation.

This paper proposes Pyramid Pooling Mamba (PPMamba), a novel network designed to address the local information loss in existing SSM-based models for RS image semantic segmentation. Compared to natural images, RS images are characterized by large variations in object scale, complex appearance changes, and extensive spatial coverage. As a result, RS tasks place higher demands on models for effective multi-scale and spatio-temporal representation learning. PPMamba consists of several layers of Pyramid Pooling-State Space Model (PP-SSM) blocks, and each block constructs multi-branch convolution-based blocks to assist the model in capturing features from each image patch. Additionally, the auxiliary multi-branch convolution-based blocks are structured in a pyramid shape in order to capture features at different scales. Since the land cover patterns in RS images are oriented in various spatial directions, the model possesses an omnidirectional state space (OSS) block to maximally establish long-distance dependencies. The structure of PP-SSM consists solely of Mamba and convolution-based blocks, leading to the capability of learning long-range dependencies with linear computational complexity. Extensive experiments on three widely used datasets, ISPRS Vaihingen, LoveDA Urban and WHU Buildings, validate the effectiveness of PPMamba. The results show that PPMamba outperforms several state-of-the-art models, highlighting its potential to address the unique challenges of RS image semantic segmentation. The main contributions of this article can be summarized as follows:

1. A novel Mamba-based network, PPMamba, is proposed to effectively model local and global relationships in RS images while maintaining linear computational complexity. By integrating CNN-based pyramid pooling and the Mamba model, PPMamba addresses the limitations of existing methods in balancing fine-grained local feature extraction with comprehensive global context modeling.

2. The core structure of PPMamba, the PP-SSM block, introduces a pyramid-shaped convolutional module combined with OSS. This block effectively fuses multi-scale local features, selectively scanned from eight different directions, with global features, enhancing the model's ability to capture diverse land cover patterns in RS images.

The remainder of this paper is organized as follows. Section 2 reviews the related works on architectures and techniques relevant to PPMamba, while Section 3 details the proposed method. Section 4 presents the experimental results and discussions, followed by the conclusion in Section 5.

<u>Notation</u>: Vectors and matrices are denoted by bold-face letters. $\boldsymbol{I}_N$ is the $N \times N$ identity matrix while $[\cdot]^T$ denotes the transposition of the enclosed vector.

2 Related Work

2.1 *Remote Sensing Image Semantic Segmentation*

Early approaches for RS image semantic segmentation primarily relied on traditional image processing techniques and classical machine learning algorithms. Methods such as pixel-level classification were widely adopted, with techniques like the Maximum Likelihood Classifier (MLC) [31] and Support Vector Machine (SVM) [28] being popular due to their simplicity and effectiveness. However, these methods typically struggled to capture spatial information and often underperformed when dealing with complex object categories in high-dimensional data.

With the emergence of deep learning, CNN and transformer-based models have demonstrated significant potential in RS image segmentation [24, 26, 32]. CNN-based models, such as ResUNet-a [7], leverage hierarchical feature extraction through convolutional layers and have been enhanced with techniques like residual connections and pyramid scene parsing. However, CNNs are limited by their local receptive fields, making it challenging to capture long-range dependencies. To address this problem, transformer-based models, such as GLOTS [22], have been introduced, utilizing self-attention mechanisms to capture global context. Despite their strengths, transformers are computationally intensive, leading to high resource demands for processing high-resolution RS images [33]. These challenges highlight the need for new architectures that balance segmentation accuracy and computational efficiency.

2.2 *Mamba*

The Mamba architecture was introduced as an alternative to transformers, addressing their high computational complexity while capturing long-range dependencies in visual data. Mamba is based on the structured state space model (SSM), originally designed to handle continuous data with linear time complexity [12]. The transition from SSM to structured state space sequence models (S4) allowed for effective processing of discrete data [12]. More specifically, we consider a continuous system that maps a 1-D function or sequence $x(t) \in \mathbb{R} \mapsto y(t) \in \mathbb{R}$ through a hidden state $\boldsymbol{h}(t) \in \mathbb{R}^{N \times 1}$. This process can be described as a linear Ordinary Differential Equation (ODE) [12]:

$$
\begin{aligned}
\boldsymbol{h}'(t) &= \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{b}x(t), \\
y(t) &= \boldsymbol{c}^T \boldsymbol{h}(t),
\end{aligned}
\tag{1}
$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ denotes the state transition matrix while $\boldsymbol{b} \in \mathbb{R}^{N \times 1}$ and $\boldsymbol{c} \in \mathbb{R}^{N \times 1}$ are the projection parameters. Furthermore, $\boldsymbol{h}'(t)$ stands for the derivative of $\boldsymbol{h}(t)$. To adapt the system to a discrete form, a zero-order hold

(ZOH) is required to convert all the parameters into their discrete counterparts, as follows:

$$
\begin{aligned}
\bar{\boldsymbol{A}} &= \exp(\triangle \boldsymbol{A}), \\
\bar{\boldsymbol{b}} &= (\triangle \boldsymbol{A})^{-1}(\exp(\triangle \boldsymbol{A}) - \boldsymbol{I}_N) \cdot \triangle \boldsymbol{b},
\end{aligned}
\tag{2}
$$

where $\triangle$ is a step size that denotes the input's resolution, $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{b}}$ are the discrete version of the projection parameters $\boldsymbol{A}$ and $\boldsymbol{b}$, respectively.

However, the S4 model faced challenges in optimizing computational efficiency, which led to the development of the selective structured state space model (S6) [11]. S6 forms the core of Mamba, introducing dynamic adjustments to $\boldsymbol{b}$, $\boldsymbol{c}$, and $\triangle$ that depend on the input, enabling hardware-aware optimizations and selective compression of information.

Recently, numerous SSM-based models have been applied across various domains, including computer vision and remote sensing. In computer vision, Vmamba and Vision Mamba have introduced innovative approaches leveraging SSM-based architectures. Vmamba maintains linear complexity while preserving global receptive fields by incorporating a Cross-Scan Module (CSM) that traverses the spatial domain and transforms non-causal visual images into ordered patch sequences [21]. Furthermore, Vision Mamba demonstrates that self-attention is not necessary for visual learning by exploiting bidirectional Mamba blocks with position embeddings to structure images and bidirectional state space models for compression [46]. In remote sensing, RS-Mamba presents an innovative architecture for image classification, introducing a dynamic multi-path activation mechanism to enhance Mambas capability in modeling non-causal data [6]. Recently, Pan-Mamba was developed to perform cross-modal information exchange by integrating channel swapping and cross-modal Mamba designs, enabling efficient fusion across modalities [15]. Additionally, Mamba-in-Mamba has shown strong performance in hyperspectral image classification [45], while ChangeMamba pioneers the application of the Mamba architecture for RS change detection tasks [4]. Despite these advancements, most of the above models are not explicitly designed for semantic segmentation. To address this challenge, RS3Mamba was proposed as one of the earliest SSM-based models tailored for RS image semantic segmentation [25]. Following this, PyramidMamba introduced an adaptable decoder featuring dense spatial pyramid pooling (DSPP) to capture multiscale semantic features [38]. However, RS3Mambas complex architecture imposes significant computational overhead, and PyramidMambas emphasis on pyramid pooling in the decoder may lead to suboptimal multiscale feature extraction within its encoder. AfaMamba [3] introduced a Mamba-based multiscale feature adaptive fusion module and a global-local Mamba block. HMAFNet [34] integrated both channel and spatial attention mechanisms into the fusion pathway of the Mamba module. While their fusion approach effectively captures spatial

context through attention mechanisms, it remains relatively static and lacks directional modeling. As a result, it struggles to precisely distinguish complex object boundaries and heterogeneous textures. MLMamba [9], on the other hand, proposed the pyramid Mamba structure and feature-guided semantic modules for multi-label remote sensing scene classification, significantly enhancing the modeling of semantic associations among labels. However, its design is primarily tailored for image-level classification tasks and lacks the capacity for pixel-level spatial modeling.

2.3   Spatial Pyramid Pooling

Spatial Pyramid Pooling (SPP) was developed to address the rigid input size requirements of early CNN architectures, enabling models to handle variable input sizes without losing critical spatial information [14]. By introducing multilevel pooling operations, SPP allows models such as AlexNet [17] and VGGNet [30] to preserve spatial hierarchies while generating fixed-length output vectors. This capability has proven essential in high-resolution image tasks, where resizing can distort important features. In RS image segmentation, SPP has been widely adopted for multiscale feature extraction, providing flexibility in adapting to the diverse spatial patterns found in RS imagery. Advanced architectures, such as Faster RCNN [10] and YOLO [29], have integrated SPP to enhance their object detection accuracy by better capturing context across different scales. Despite these advances, current models often emphasize either local detail (as in CNN-based approaches) or global context (as in transformer-based models), leading to suboptimal performance in scenarios requiring a nuanced understanding of both. The challenge remains to develop an architecture that effectively integrates multiscale local and global features while maintaining computational efficiency.

3   Methodology

3.1   Proposed PPMamba

The proposed PPMamba architecture is illustrated in Figure 1. The input to the model is an image with dimensions $H \times W \times 3$, processed through a UNet-like encoder-decoder framework. The encoder reduces the spatial resolution of the input while preserving essential features. Furthermore, the decoder progressively upsamples the features to produce the final segmentation map. In the encoder, the input image first undergoes a patch embedding operation, converting it into feature maps of size $\frac{H}{4} \times \frac{W}{4} \times C$. These feature maps are then passed through a sequence of patch merging operations and PP-SSM blocks. The patch merging operations successively reduce the spatial resolution from
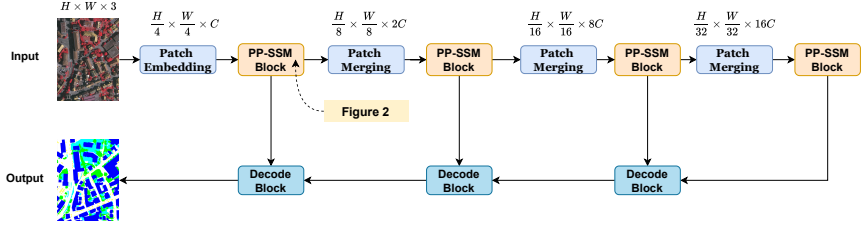
Figure 1: The architecture of proposed PPMamba.

$\frac{H}{4} \times \frac{W}{4}$ to $\frac{H}{32} \times \frac{W}{32}$, while increasing the number of channels to $16C$. The stacked PP-SSM blocks enable the model to capture both local and global context information while maintaining computational efficiency. The decoder consists of four stages of upsampling. Each decoder block fuses the upsampled features with the corresponding encoder features and features from its previous decode block, enabling the reconstruction of detailed spatial information. The output is a high-resolution segmentation map with dimensions $H \times W \times 3$.

3.2   Proposed PP-SSM Block

Figure 2(a) shows the structure of a conventional visual SSM block in which input is processed by visual state space (VSS) blocks followed by a layer normalization (LN) block and a multilayer perceptron (MLP) block. However, the VSS block suffers from many limitations in capturing global spatial features from RS images.

In sharp contrast, the proposed PP-SSM block, shown in Figure 2(b), is the core structure in our PPMamba model, utilizing a multi-branch auxiliary methodology for RS image semantic segmentation. First, the input is separated into four distinct parts along the channel dimension, namely $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ and $\boldsymbol{x}_4$, as shown in Figure 2(b). This separation allows the PP-SSM block to independently capture different aspects of the local features using four SPP branches. These SPP branches stack continuous convolutional layers with different kernel sizes to capture the local features while maintaining the input's resolution the same way as the output to preserve the local spatial information. Specifically, $\boldsymbol{x}_2, \boldsymbol{x}_3$ and $\boldsymbol{x}_4$ are passed through two layers of convolutional blocks of kernel sizes 3×3, 5×5 and 7×7 respectively, followed by ReLU activation functions. Finally, the resulting features together with $\boldsymbol{x}_1$ are processed by 2D convolutional blocks of kernel size 1×1.

It is worth pointing out that employing various kernel sizes to process $\boldsymbol{x}_2, \boldsymbol{x}_3$ and $\boldsymbol{x}_4$ can form a pyramid structure, enabling the model to capture a wider range of local features at different scales. The pyramid-shaped design
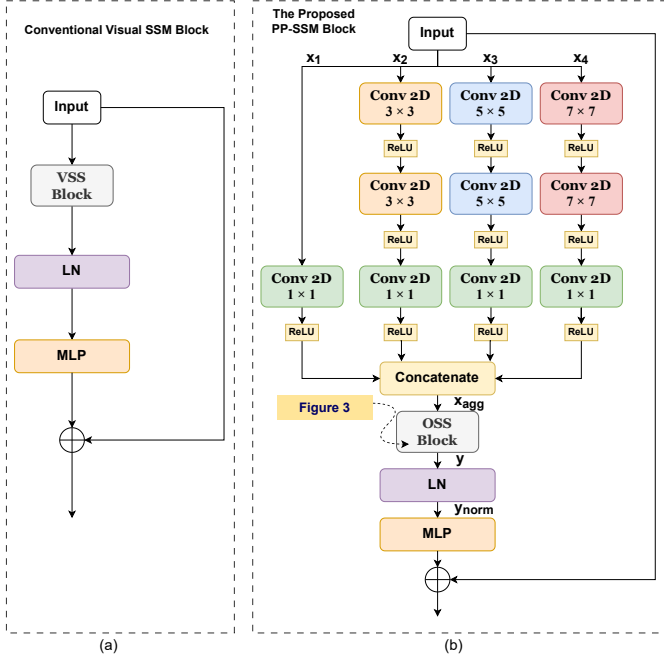
Figure 2: The architectures of the conventional visual SSM block and the proposed PP-SSM block. (a) The architecture of a conventional visual SSM block. (b) Proposed architecture of PP-SSM Block.

is crucial for extracting comprehensive local features from the input image, which is essential for accurate semantic segmentation. The output of each convolution-based block will be passed through a ReLU activation that introduces non-linearity into the model and enhances its capability to learn complex patterns from the input data.

After processing through the convolutional layers, the PP-SSM block concatenates the outputs to form a unified feature map x_{agg} with the same number of channels as the original input. After that, x_{agg} is fed into an omnidirectional state space block (OSS) [43] to capture the global features of the RS images. The OSS block performs selective scanning in multiple directions to capture global dependencies and spatial relationships from various angles. Detailed operations of the OSS will be elaborated in the next section. The output of OSS, denoted as y, is first normalized before being processed by an MLP block. The normalization block makes the training process converge faster, while the MLP block can adjust the input dimensions.

In summary, the PP-SSM block introduces four convolution-based branches with various kernel sizes to collect local features. Furthermore, the pyramid-shaped kernel sizes capture features across different dimensions.

3.3 Omnidirectional State Space Block (OSS)

As shown in Figure 3(a), the architecture of the proposed OSS block begins with a layer normalization stage to stabilize the training process. Next, a linear transformation adjusts the input dimensions before the data passes through a depthwise convolution operation (DWConv) to extract spatial features. The core structure of the OSS block, called the omnidirectional selective scan module (OSSM), then selectively scans the features in forward and backward directions across four different angles, i.e., eight scanning directions, as depicted in Figure 3(b). Finally, the output is passed through a linear transformation before the residual connections are applied to concatenate the input features with the final output.
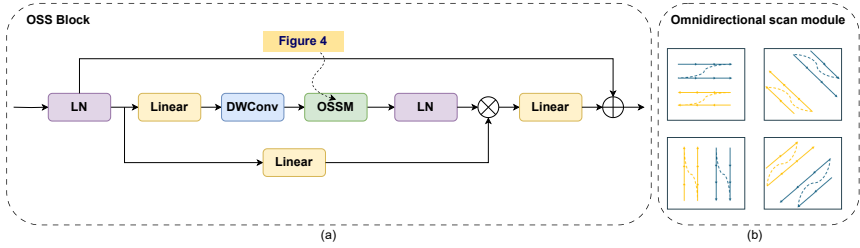

Figure 3: (a) The architecture of the proposed OSS block. (b) The illustration of the selective scan directions of OSSM.

The operation of OSSM is illustrated in Figure 4. We denote by φ_{in} and φ_{out} the input and output features of OSSM, respectively. The scanning process can be described as follows:

$$
\begin{aligned}
\varphi_{in}^n &= expand(\varphi_{in}, n), \\
\varphi_{in}^{\bar{n}} &= w_{((n-1)//2)+1} \cdot S6(\varphi_{in}^n), \\
\varphi_{out} &= merge(\varphi_{in}^{\bar{1}}, \varphi_{in}^{\bar{2}}, \varphi_{in}^{\bar{3}}, \varphi_{in}^{\bar{4}}, \varphi_{in}^{\bar{5}}, \varphi_{in}^{\bar{6}}, \varphi_{in}^{\bar{7}}, \varphi_{in}^{\bar{8}}),
\end{aligned}
\tag{3}
$$

where $n \in N = \{1, 2, 3, ..., 8\}$ represents the eight different scanning directions. Furthermore, $expand(\cdot)$ and $merge(\cdot)$ denote the scan expansion and merging operation, respectively. $S6(\cdot)$ is the selective scan space state sequential model [11]. There are four adaptive weights w_1, w_2, w_3, w_4. w_1 represents the significance of "Z" shape and anti "Z" shape scanning methods. w_2 denotes the significance of "N" shape and anti "N" shape. $w_3 and w_4$ indicate the
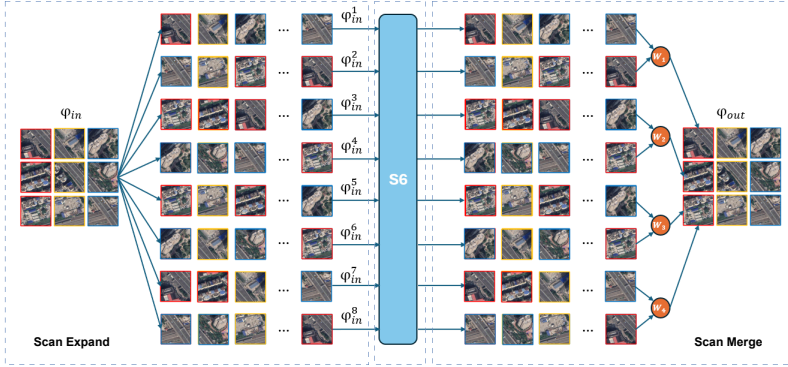
Figure 4: Illustration of the operation of the proposed oriented scanning module (OSSM).

significance of diagonal and anti-diagonal scanning methods. Since different scanning methods may vary from shallow features to deep features, the values of weights will change to adapt to different levels of images.

4 Experiments

4.1 Datasets

4.1.1 ISPRS Vaihingen

The Vaihingen dataset consists of high-resolution aerial images captured over Vaihingen, Germany, as part of the German Association of Photogrammetry and Remote Sensing (DGPF) benchmark. The dataset contains 16 true orthophotos, each of resolution 2500×2000 pixels. For our experiments, 12 orthophotos were used as the training set, and the remaining 4 orthophotos were used for testing. The training set includes images with indices $1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34$, and 37, while the test set comprises images with indices $5, 21, 15$, and 30. Each orthophoto contains three spectral bands: near-infrared (NIR), red, and green (NIRRG). The ground sampling distance is 9 centimeters, and the dataset is annotated with five foreground classes: impervious surfaces, buildings, low vegetation, trees, and cars, along with a background class.

4.1.2 LoveDA Urban

The LoveDA dataset [36, 37] provides high-resolution RS images, with 5987 samples in total, captured over three cities in China: Nanjing, Changzhou, and Wuhan. For this study, we focus on the urban subset, which includes

1833 images, each with a resolution of 1024×1024 pixels. The dataset is split into 1156 training images and 677 testing images. The training set consists of images indexed from No. 1366 to No. 2521, while the test set covers indices from No. 3514 to No. 4190. The images are provided in three channels: red, green, and blue (RGB), with a ground sampling distance of 30 centimeters. The LoveDA Urban dataset includes seven land cover classes: background, buildings, roads, water, barren land, forests, and agriculture.

4.1.3 WHU Building

The WHU Building dataset [16] originates from aerial imagery provided by the New Zealand Land Information Services. To prepare the data for semantic segmentation tasks, most of the aerial images were downsampled to a ground resolution of 30 centimeters and subsequently cropped into 8188 tiles, each with a size of 512×512 pixels. The resulting ready-to-use dataset is partitioned into three subsets: a training set containing 5732 images, a validation set with 1228 images, and a test set including 1228 images. The images are provided in three channels: red, green, and blue (RGB) with two categories: background and buildings.

4.2 Evaluation Metrics

Mean intersection over union (mIoU) and mean F1-score (mF1) were used to evaluate the performance of the models. Besides, *precision* and *recall* were used to calculate the F1-score. The definitions and equations for these metrics are as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

$$F1 - score = \frac{2(Precision \cdot Recall)}{Precision + Recall},$$

$$mF1 = \frac{1}{k+1} \sum_{i=0}^{k} \frac{2(Precision \cdot Recall)}{Precision + Recall}, \tag{6}$$

$$IoU = \frac{TP}{FN + FP + TP},$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP}, \tag{7}$$

where k is the number of categories, TP denotes true positives, FP denotes false positives, and FN denotes false negatives.

4.3   Loss Functions

In traditional semantic segmentation tasks, the loss function is typically based on the cross-entropy-based segmentation loss:

$$Loss_{seg} = -\sum_{H,W}\sum_{c \in C} Y \log P, \tag{8}$$

where $Y \in \mathbb{R}^{H \times W \times C}$ and $P \in \mathbb{R}^{H \times W \times C}$ denote the ground truth and predicted output, respectively. C represents the categories of the ground objects of the RS images.

To better preserve the continuity of pixels within objects and introduce edge constraints, which are both important for improving segmentation accuracy, we adopt a composite loss function [23] defined as:

$$Loss_{total} = Loss_{seg} + \lambda_o Loss_{obj} + \lambda_b Loss_{bdy}, \tag{9}$$

where $Loss_{obj}$ is the object consistency loss and $Loss_{bdy}$ is the boundary preservation loss [2, 23]. λ_o and λ_b are weighting coefficients. This integrated loss design helps guide the model toward more precise and structurally coherent predictions during training.

4.4   Implementation Details

Stochastic gradient descent (SGD) was applied as the optimization algorithm for training all models. The learning rate, momentum, and decaying coefficient values were set to $0.01, 0.9$, and 0.0005, respectively. The batch size was set to 10, while the epoch size was set to 50. The number of PP-SSM blocks at each stage is $[2, 2, 9, 2]$. No pre-trained strategy is loaded in order to confirm the effectiveness of the PPMamba architecture. Evaluation metrics were calculated twice per epoch. The experiments were conducted on a server node running Ubuntu 22.04.1 operating system, equipped with an NVIDIA GeForce RTX 4090 GPU. The framework utilized in these experiments was PyTorch 2.2.2.

4.5   Performance Comparison

To evaluate the effectiveness of PPMamba, we conducted comparative experiments against nine state-of-the-art models. The baseline model used in these experiments is RS-Mamba [43]. The comparison models include CNN-based methods, ABCNet [20], MANet [19], and CMTFNet [41], transformer-based methods, FTUNetFormer [39], hybrid CNN-transformer models, UNetFormer [39], HST_UNet [44], and TransUNet [5], and other Mamba-based models, RS3Mamba [25].

4.5.1 Performance Comparison on ISPRS Vaihingen

As shown in Table 1, PPMamba demonstrated significant improvement over its baseline model, RS-Mamba. The primary evaluation metrics, mIoU and mF1, increased by 6.85% and 4.62%, respectively, confirming that RS-Mamba has limitations in RS image semantic segmentation tasks, which PPMamba effectively overcame. Notably, PPMamba achieved the best performance across all five foreground classes except for tree. For the impervious surface class, PPMamba achieved an F1 score of 92.20%, nearly 1.30% higher than RS3Mamba, underscoring its ability to distinguish between urban structures and other land cover types. It also led in the building class, surpassing the baseline model by 3.72% on F1 score. This superior performance suggests that PPMamba excels at capturing complex building shapes and boundaries, which are often challenging due to occlusions and shadows. In the low vegetation class, PPMamba outperformed ABCNet by 1.22% and FTUNetFormer by 2.67% on F1 score, highlighting its accuracy in identifying and segmenting areas covered by grass, shrubs, and other low-height vegetation. Furthermore, PPMamba achieved the highest F1 score (83.86%) and IoU (80.10%) in the car category, surpassing RS-Mamba by 11.66% and 15.86%. This improvement reflects its enhanced ability to recognize local features, especially when detecting cars, occupying only a small portion of the Vaihingen images. These results have demonstrated the potential of PPMamba in effectively recognizing a wide range of categories.

Table 1: The performance of PPMamba and other state-of-the-art models on the Vaihingen dataset, where type C indicates CNN-based models, T indicates transformer-based models, C-T indicates CNN&transformer-based models, and M indicates Mamba-based models. F1-score and IoU are chosen as evaluation metrics. The accuracy of each category is presented by F1/IoU. **Bold** font represents the best values.

Model	Type	impervious surface	building	low vegetation	tree	car	mF1	mIoU
ABCNet [20]	C	89.68/90.45	93.72/93.90	77.93/75.52	89.81/91.07	73.46/63.16	84.92	74.57
MANet [19]	C	90.28/91.74	94.28/93.07	78.95/**79.26**	89.85/89.76	77.58/70.76	86.19	76.32
CMTFNet [41]	C	90.69/90.50	95.03/96.20	78.89/76.18	90.13/91.33	82.09/74.95	87.37	78.06
FTUNetFormer [39]	T	90.78/90.37	94.54/94.88	76.48/73.59	89.15/91.83	75.28/66.49	85.25	75.09
UNetFormer [39]	C-T	90.37/**92.19**	94.58/93.44	78.37/76.56	90.19/91.15	81.85/75.87	87.07	77.60
HST_UNet [44]	C-T	91.27/91.34	95.36/95.43	78.44/77.27	90.04/91.02	83.61/79.07	86.62	78.67
TransUNet [5]	C-T	91.24/90.31	94.82/96.63	78.85/74.71	**90.54/92.79**	83.77/78.97	87.84	78.78
RS3Mamba [25]	M	90.87/89.99	95.26/95.59	78.49/75.74	90.20/91.93	81.83/74.10	87.33	78.04
RS-Mamba [43]	M	88.37/87.73	92.52/92.08	76.31/75.68	89.14/90.14	72.20/64.24	83.71	72.77
PPMamba	M	**92.20**/91.40	**96.24/96.92**	**79.15**/76.49	90.19/92.43	**83.86/80.10**	**88.33**	**79.62**

Figure 5 presents a visual comparison of segmentation results on the ISPRS Vaihingen dataset, including outputs from all models, the NIRRG image, and the ground truth. The visual results have shown that PPMamba provided more accurate and detailed segmentation, particularly in building boundaries, tree and low vegetation regions. Notably, PPMamba correctly identified the cars in the parking lot of the building in the left part of the image. Additionally, PPMambas segmentation of buildings (blue areas) maintained continu-
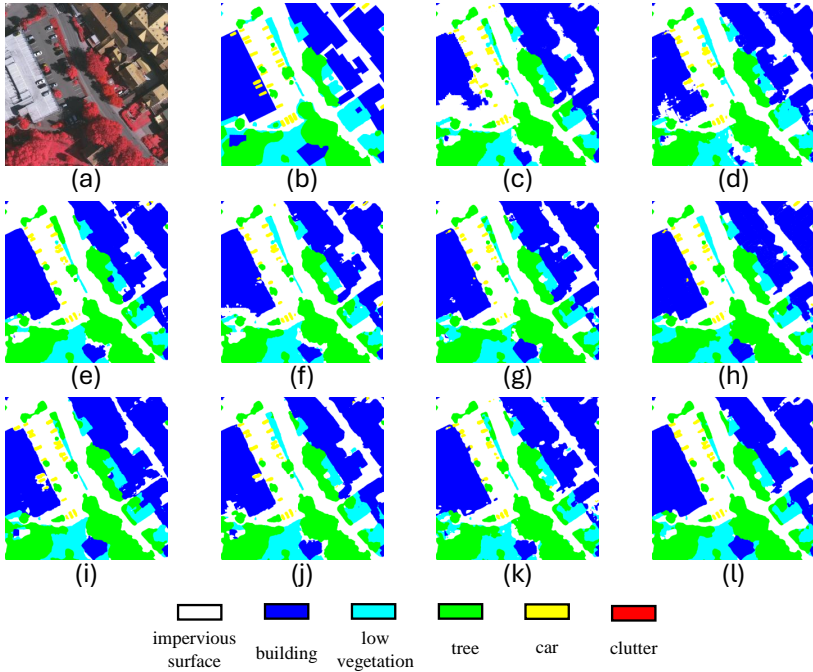
Figure 5: Performance comparisons on the ISPRS Vaihingen dataset with the size of 1024×1024. (a) NIRRG images, (b) Ground truth, (c) ABCNet, (d) MANet, (e) CMTFNet, (f) UNetFormer, (g) FTUNetFormer, (h) HST_UNet, (i) TransUNet, (j) RS3Mamba, (k) RS-Mamba and (l) PPMamba.

ous and precise outlines at the bottom of the image, with building boundaries seamlessly connecting to those of trees and low vegetation without any gaps. In contrast, the blue areas produced by other comparison models, including our baseline model, RS-Mamba, showed blurred and jagged edges. PPMamba also excelled in distinguishing between low vegetation and tree classes, where other models often suffered from over-segmentation.

4.5.2 Performance Comparison on LoveDA Urban

As shown in Table 2, experiments have been performed on the LoveDA Urban dataset to further validate the performance of PPMamba. Similar to the results on the previous dataset, PPMamba achieved the highest mIoU and mF1 scores among the nine state-of-the-art models. It significantly outperformed the baseline model, with an improvement of 8.69% in mIoU and 8.47% in mF1, due to its superior capability in capturing local features in RS images

Table 2: The performance of PPMamba and other state-of-the-art models on the LoveDA Urban dataset, where type C indicates CNN-based models, T indicates transformer-based models, C-T indicates CNN&transformer-based models, and M indicates Mamba-based models. F1-score and IoU are chosen as evaluation metrics. The accuracy of each category is presented by F1/IoU. **Bold** font represents the best values.

Model	Type	background	building	road	water	barren	forest	agriculture	mF1	mIoU
ABCNet [20]	C	51.79/66.54	67.87/70.66	64.94/56.57	67.56/58.12	45.94/35.23	54.02/77.60	**28.93**/19.09	59.62	43.05
MANet [19]	C	51.42/63.92	70.55/**74.98**	65.37/64.73	70.17/64.60	**48.33/40.67**	52.69/84.92	5.22/2.81	61.17	44.72
CMTFNet [41]	C	52.57/67.64	70.05/74.06	68.81/60.70	69.09/58.19	37.70/25.77	54.29/83.01	26.18/16.59	59.64	43.60
FTUNetFormer [39]	T	49.84/61.82	69.91/70.42	68.88/64.03	67.73/56.98	26.81/16.82	51.40/89.96	23.29/14.18	56.63	41.23
UNetFormer [39]	C-T	51.57/64.48	69.10/70.94	64.40/61.48	67.07/**65.24**	44.49/34.20	54.23/82.87	16.99/9.95	59.32	42.82
HST_UNet [44]	C-T	50.39/**68.61**	71.07/70.02	70.57/61.95	68.67/58.01	17.67/10.01	54.00/84.64	27.76/16.70	55.68	41.06
TransUNet [5]	C-T	52.47/67.23	67.19/60.93	67.03/59.46	**73.68**/63.13	40.59/31.84	43.75/85.72	0.00/0.00	60.19	44.07
RS3Mamba [25]	M	51.03/67.73	69.98/70.53	68.86/63.39	70.51/61.57	41.52/28.46	**58.20**/85.67	23.59/14.50	60.38	44.25
RS-Mamba [43]	M	48.83/64.63	60.62/55.34	59.64/53.95	67.92/54.16	35.36/26.26	47.85/**90.74**	3.51/1.80	54.47	38.24
PPMamba	M	**54.94**/68.03	**71.82**/71.86	**73.28/67.41**	70.91/63.29	43.75/31.70	52.83/88.17	24.87/15.11	**62.94**	**46.93**

compared to the baseline model, RS-Mamba. In particular, PPMamba exhibited impressive performance in the background, building and road categories. Specifically, PPMamba achieved an IoU of 68.03% and an F1 score of 54.94% in the background class, ranking among the best performers across all models. This highlights PPMambas ability to accurately capture background features and effectively distinguish them from adjacent categories. In the building class, our model also achieved the highest F1 scores (71.82%), demonstrating its strength in precisely segmenting building structures. For the road category, PPMamba attained the top F1 score of 73.28%, outperforming the second-best model, HST_UNet, by 2.71%, and it also led in IoU with a score of 67.41%. These results underscored PPMambas excellent ability to segment road areas. Additionally, PPMamba outperformed the baseline model, RS-Mamba, by 21.36% in F1 score for the agriculture category, indicating notable improvement in recognizing agricultural areas.

Figure 6 provides a visual comparison of the test results across all models, along with the RGB image and the ground truth. In the top-right corner, a square red area is clearly delineated by PPMamba, which accurately captured the square building with clear and contiguous outlines, free from significant errors. In contrast, the baseline model RS-Mamba and other state-of-the-art models such as CMTFNet and UNetFormer struggled with this task. They failed to clearly outline the square shape, with UNetFormer even misclassifying parts of the building as roads. RS-Mamba also had difficulty in detecting the yellow road area in the lower part of the image, leading to blurred boundaries between the road and building classes. This resulted in some road areas being incorrectly classified as buildings (red). In contrast, PPMamba produced continuous and precise boundaries for road areas, clearly distinguishing them from adjacent classes.
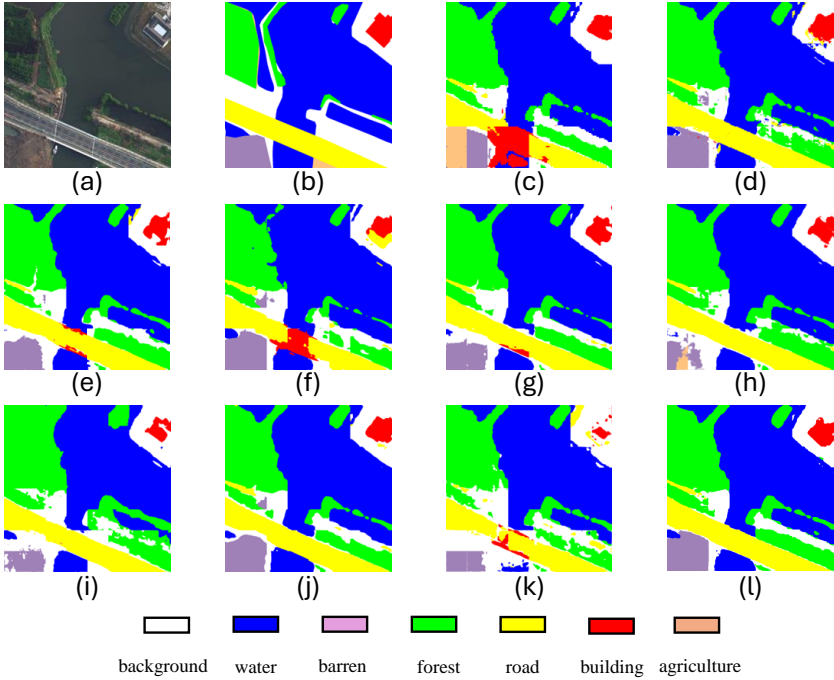
Figure 6: Performance comparisons on the LoveDA Urban dataset with the size of 1024 × 1024. (a) RGB images, (b) Ground truth, (c) ABCNet, (d) MANet, (e) CMTFNet, (f) UNetFormer, (g) FTUNetFormer, (h) HST_UNet, (i) TransUNet, (j) RS3Mamba, (k) RS-Mamba and (l) PPMamba.

4.5.3   Performance Comparison on WHU Building

The WHU Building dataset is employed as a supplementary benchmark to demonstrate that PPMamba maintains leading performance on non-traditional RS semantic segmentation tasks, thereby enhancing the robustness and generalizability of the proposed method. Figure 7 shows that the RS images processed by PPMamba have clearer and more rectilinear building contours. Table 3 highlights the top two values for each metric in bold. On the WHU Building dataset, TransUNet achieved the highest overall mIoU (92.05%) and mF1 (95.78%). However, its FLOPs and parameter count are nearly five times greater than those of PPMamba. Given the substantial computational complexity of TransUNet, a trade-off between performance and efficiency becomes necessary. PPMamba emerges as a more balanced choice, achieving the second-best mIoU (91.77%) and mF1 (95.62%), both closely approaching TransUNet's results. More importantly, PPMamba achieves the
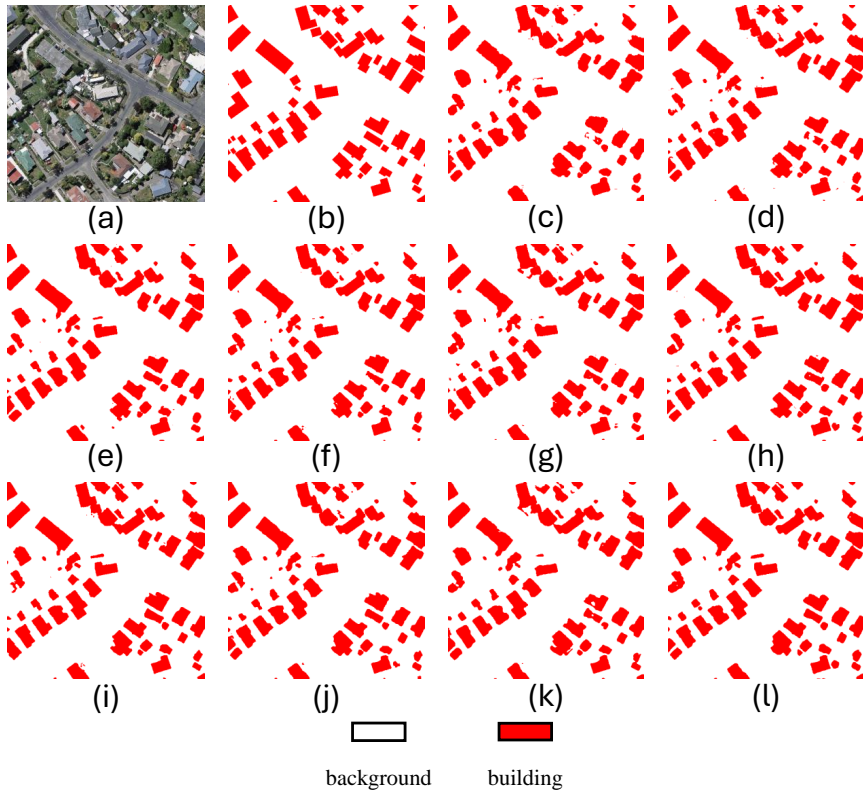
Figure 7: Performance comparisons on the WHU Building dataset with the size of 512 × 512. (a) RGB images, (b) Ground truth, (c) ABCNet, (d) MANet, (e) CMTFNet, (f) UNetFormer, (g) FTUNetFormer, (h) HST_UNet, (i) TransUNet, (j) RS3Mamba, (k) RS-Mamba and (l) PPMamba.

highest IoU (91.76%) for the building category, underscoring its strong capability in recognizing specific geographical features.

In summary, the comparative results across three different datasets have demonstrated the significant potential of PPMamba in RS image semantic segmentation, which confirms that PPMamba is more competitive and effective than both its baseline model and other state-of-the-art models mentioned in this study.

4.6   Feature Capture Capability Comparison

Mamba excels at capturing long-range dependencies [11], but its ability to extract local features is less effective. This experiment aims to analyze the

Table 3: The performance of PPMamba and other state-of-the-art models on the WHU Building dataset, where type C indicates CNN-based models, T indicates transformer-based models, C-T indicates CNN&transformer-based models, and M indicates Mamba-based models. F1-score and IoU are chosen as evaluation metrics. The accuracy of each category is presented by F1/IoU. **Bold** font represents the top 2 values.

Model	Type	background	building	mF1	mIoU
ABCNet [20]	C	98.33/98.73	90.68/88.68	94.51	89.84
MANet [19]	C	98.53/98.85	91.84/90.22	95.19	91.01
CMTFNet [41]	C	98.58/98.82	92.14/90.91	95.36	91.32
FTUNetFormer [39]	T	98.46/98.77	91.45/89.89	94.95	90.60
UNetFormer [39]	C-T	98.50/98.79	91.70/90.26	95.10	90.86
HST_UNet [44]	C-T	98.64/**98.87**	92.50/91.33	95.57	91.68
TransUNet [5]	C-T	**98.71/98.95**	**92.85/91.62**	**95.78**	**92.05**
RS3Mamba [25]	M	98.51/98.83	91.70/90.07	95.10	90.86
RS-Mamba [43]	M	98.25/98.58	90.26/88.59	94.26	89.40
PPMamba	M	**98.65**/98.81	**92.58/91.76**	**95.62**	**91.77**

differences in local feature extraction between the baseline model RS-Mamba and our enhanced model PPMamba using heatmaps. In Figure 8, the category of the red pixel at coordinates [99, 49] is labeled as "buildings" in subimage (a) and "impervious surfaces" in subimage (b). In these heatmaps, red indicates a higher likelihood of predicting the designated category, while blue suggests little to no correlation. In the last two rows of both (a) and (b) in Figure 8, the feature maps with sizes $[1, 768, 8, 8]$, $[1, 384, 16, 16]$, and $[1, 192, 32, 32]$ are shown in the format $[B, C, H, W]$ from left to right in each row. The NIRRG images were taken from the ISPRS Vaihingen dataset, where a 256×256 window was slid across the images with a set stride, generating the NIRRG images in the heatmaps.

Figure 8 compares the feature extraction capabilities of RS-Mamba and PPMamba in two selected scenarios. In subimages (a_3)-(a_5), RS-Mamba frequently misclassified buildings and nearby low vegetation or impervious surfaces as similar features. As a result, large patches of red and yellow were scattered across the subimage (a_4). In contrast, PPMamba demonstrated superior local feature extraction. In the last two subimages (a_7) and (a_8), PPMamba delineated the contours of all buildings, highlighting them with prominent red and yellow regions that closely aligned with the ground truth. In scenario (b), RS-Mamba struggled to differentiate between impervious surfaces and buildings. In subimage (b_5), red and green regions erroneously covered the building category. On the other hand, PPMamba exhibited better performance in subimages (b_6)-(b_8), recognizing the shape of impervious surfaces not only in the low-level feature map but also in the high-level one. Table 1 further supports this analysis, showing that PPMamba achieved the highest F1 scores for both buildings (96.24%) and impervious surfaces (92.20%) among all state-of-the-art models. These heatmap comparisons have clearly
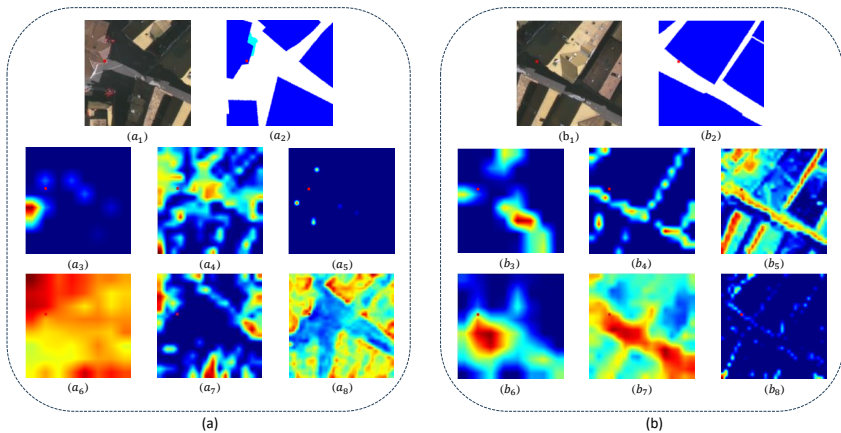
Figure 8: The comparison of heatmaps of RS-Mamba and PPMamba. $(a)(a_1)$ the NIRRG image, (a_2) the ground truth, $(a_3$-$a_5)$ three heatmaps from RS-Mamba, $(a_6$-$a_8)$ three heatmaps from PPMamba. (b) are organized in the same way. (a) Heatmaps of encoders determine if a pixel belongs to *buildings* or not. (b) Heatmaps of encoders determine if a pixel belongs to *impervious surfaces* or not. The selected NIRRG images and ground truth are slid with a fixed window size from the ISPRS Vaihingen dataset.

demonstrated that PPMamba offers a more effective local feature extraction capability than its baseline model, RS-Mamba.

4.7 Ablation Study

To validate the effectiveness of the proposed multi-branch auxiliary architecture, pyramid-shaped convolutional blocks, adaptive weights and OSS block, eight ablation experiments were conducted on both the ISPRS Vaihingen and LoveDA Urban datasets. In Table 4, the first row for each dataset represents the baseline model RS-Mamba, which does not include the multi-branch convolutional auxiliary architecture. The second row corresponds to a version of PPMamba with four convolutional branches, but with all branches having identical 1×1 kernel sizes. The third row represents the full PPMamba model without adaptive weights, which combines the multi-branch auxiliary structure with pyramid-shaped kernel sizes for the convolutional blocks. The final row displays the result of multi-branch pyramid-shaped PPMamba with adaptive weights. Table 5 presents the results of the two Mamba models incorporating VSS and OSS blocks on the two datasets, respectively.

Table 4 presents the performance comparison across all four configurations. PPMamba with four identical branches showed significant improvements in evaluation metrics, increasing mIoU by 5.61% for Vaihingen and by 4.74% for Urban, and mF1 by 3.84% for Vaihingen and by 3.91% for Urban. These sub-

Table 4: The ablation study of PPMamba on ISPRS Vaihingen and LoveDA Urban dataset. **Bold** font represents the best values.

Dataset	Model	MB[1]	PS[2]	AW[3]	mF1	mIoU
Vaihingen	RS-Mamba				83.71	72.77
Vaihingen	PPMamba	$\checkmark$			87.55	78.38
Vaihingen	PPMamba	$\checkmark$	$\checkmark$		87.80	78.81
Vaihingen	PPMamba	$\checkmark$	$\checkmark$	$\checkmark$	**88.33**	**79.62**
Urban	RS-Mamba				54.47	38.24
Urban	PPMamba	$\checkmark$			58.38	42.98
Urban	PPMamba	$\checkmark$	$\checkmark$		61.76	46.14
Urban	PPMamba	$\checkmark$	$\checkmark$	$\checkmark$	**62.94**	**46.93**

[1] MB: Multi-Branch
[2] PS: Pyramid-shaped
[3] AW: Adaptive Weights

Table 5: The ablation study of OSS block on ISPRS Vaihingen and LoveDA Urban dataset. **Bold** font represents the best values.

Dataset	Model	VSS[1]	OSS[2]	mF1	mIoU
Vaihingen	PPMamba	$\checkmark$		88.08	79.12
Vaihingen	PPMamba		$\checkmark$	**88.33**	**79.62**
Urban	PPMamba	$\checkmark$		60.34	45.00
Urban	PPMamba		$\checkmark$	**62.94**	**46.93**

[1] VSS: normal visual state space block
[2] OSS: omnidirectional state space block

stantial enhancements indicate that introducing a multi-branch convolutional structure significantly strengthened RS-Mambas feature extraction capability. Furthermore, by employing varying kernel sizes of 1×1, 3×3, 5×5, and 7×7 as part of the pyramid pooling operation, PPMamba can capture local features at different scales in RS images. This resulted in further increases in mIoU and mF1 by 0.43% for Vaihingen (3.16% for Urban), and 0.25% for Vaihingen (3.38% for Urban), respectively. Moreover, after adding adaptive weights to PPMamba, the mIoUs of Vaihingen and Urban are enhanced to 79.62% and 46.93%, respectively. We further compare the Mamba models employing VSS (four-direction scanning) and OSS (eight-direction scanning) blocks. The OSS module improves the mIoU of PPMamba by 0.5% on the Vaihingen dataset and by 1.93% on the Urban dataset, clearly demonstrating its significant contribution to performance gains. Overall, the combination of the four-branch auxiliary architecture, pyramid-shaped convolutional blocks,

adaptive weights and OSS block has made PPMamba highly effective and competitive in the semantic segmentation of RS images.

4.8 Model Complexity Analysis

Table 6 presents the computational complexity analysis for all the models discussed in this paper. FLOPs, parameters, memory usage and FPS are used to comprehensively assess the complexity of PPMamba compared to other state-of-the-art models. FLOPs refers to the number of floating-point operations required to run a network model, indicating the computational load during inference. Parameters represent the number of model parameters that need to be learned, serving as an important measure of model complexity. Generally, models with more parameters have greater expressive power. Memory usage, which refers to GPU memory consumption, is influenced by both model size and batch size. In this analysis, the batch size is fixed at two, so only model size affects GPU memory usage. FPS refers to frames per second, which indicates the number of images processed per second. A higher FPS value implies faster model inference speed.

Table 6: The computational complexity analysis. FLOPs and parameter were evaluated by a random tensor with size [1, 3, 256, 256]. Memory was evaluated by NVIDIA-SMI when running the process with batch size = 2. **Bold** font represents the best values.

Model	FLOPs (G)	Parameter (M)	Memory (MiB)	FPS (img/s)	mIoU (%)
ABCNet	3.91	13.39	**1068**	**237.76**	74.57
MANet	19.45	35.86	1744	141.02	76.32
CMTFNet	8.57	30.07	1728	100.77	78.06
UNetFormer	**2.94**	**11.68**	1074	194.79	77.60
FTUNetFormer	25.51	75.16	3156	44.16	75.09
HST_UNet	11.51	29.39	1926	41.93	78.67
TransUNet	88.29	311.23	5744	58.76	78.78
RS3Mamba	15.83	49.66	2204	53.92	78.04
RS-Mamba	9.45	40.73	2698	53.14	72.77
PPMamba	15.49	68.38	3280	26.53	**79.62**

From Table 6, several insights into the complexity of PPMamba can be drawn. Firstly, PPMamba requires 15.49 G FLOPs, making it quite competitive among the selected models. This indicates that the time complexity of PPMamba is comparable to that of some CNN-based models, owing to the fast inference speed characteristic of the Mamba architecture. This advantage allows PPMamba to outperform many transformer-based models in terms of computational efficiency. In terms of parameters, PPMambas count is slightly higher at 68.38 M, primarily due to the local auxiliary mechanism, which uses a four-branch pyramid-shaped structure. The pyramid-shaped convolutional

blocks are designed to capture local features at multiple scales, adding to the models complexity. While its parameter count is slightly higher than that of MANet (35.86 M) and CMTFNet (30.07 M), it remains significantly lower than models such as FTUNetFormer (75.16 M) and TransUNet (311.93 M). Given its superior mIoU performance, PPMamba is considered an excellent choice for RS image semantic segmentation tasks. Although the multi-branch structure introduced in our model may result in a slower inference speed, it offers a lower computational complexity compared to transformer-based models while achieving superior performance. This trade-off highlights a promising research direction. In future work, we plan to explore more efficient Mamba-based architectures to further optimize both speed and accuracy.

5   Conclusion

This work has proposed a novel model called PPMamba, which integrates CNN and Mamba to address RS image semantic segmentation tasks. To mitigate the issue of local information loss, the core architecture of PPMamba, the PP-SSM block, is proposed and incorporated into the encoder. Endowed with the OSS model, the proposed PP-SSM block selectively scans feature maps in eight different directions, with a pyramid-shaped convolutional auxiliary mechanism to extract both local and global features from input images. This innovative design allows PPMamba to achieve competitive performance while maintaining linear computational complexity. To validate the effectiveness of PPMambas architecture, comprehensive experiments were conducted on three widely used RS datasets, ISPRS Vaihingen, LoveDA Urban and WHU Buildings. The results have confirmed that the proposed semantic segmentation model can substantially outperform conventional models.

References

[1] S. Bhan and K. Krishnanunni, "Applications of remote sensing techniques to geology", *Proceedings of the Indian Academy of Sciences Section C: Engineering Sciences*, 6, 1983, 297–311.

[2] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation", in *International symposium on neural networks*, Springer, 2019, 388–401.

[3] H. Chen, H. Luo, and C. Wang, "AfaMamba: Adaptive Feature Aggregation With Visual State Space Model for Remote Sensing Images Semantic Segmentation", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 2025, 8965–83, DOI: 10.1109/JSTARS.2025.3552942.

[4] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, "ChangeMamba: Remote Sensing Change Detection With Spatiotemporal State Space Model", *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024, 1–20, DOI: 10.1109/TGRS.2024.3417253.

[5] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, *et al.*, "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers", *Medical Image Analysis*, 2024, 103280.

[6] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, "RSMamba: Remote Sensing Image Classification With State Space Model", *IEEE Geoscience and Remote Sensing Letters*, 21, 2024, 1–5, DOI: 10.1109/LGRS.2024.3407111.

[7] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data", *ISPRS Journal of Photogrammetry and Remote Sensing*, 162, 2020, 94–114.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", in *International Conference on Learning Representations*, 2021, https://openreview.net/forum?id=YicbFdNTTy.

[9] R. Du, X. Tang, J. Ma, X. Zhang, and L. Jiao, "MLMamba: A Mamba-based Efficient Network for Multi-label Remote Sensing Scene Classification", *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, 1–1, DOI: 10.1109/TCSVT.2025.3535939.

[10] R. Girshick, "Fast R-CNN", in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, 1440–8, DOI: 10.1109/ICCV.2015.169.

[11] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces", *arXiv preprint arXiv:2312.00752*, 2023.

[12] A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces", in *The International Conference on Learning Representations (ICLR)*, 2022.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–8.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", *IEEE transactions on pattern analysis and machine intelligence*, 37(9), 2015, 1904–16.

[15] X. He, K. Cao, K. Yan, R. Li, C. Xie, J. Zhang, and M. Zhou, "Pan-Mamba: Effective pan-sharpening with State Space Model", *arXiv preprint arXiv:2402.12192*, 2024.

[16] S. Ji, S. Wei, and M. Lu, "Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set", *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 2019, 574–86, DOI: 10.1109/TGRS.2018.2858817.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in neural information processing systems*, 25, 2012.

[18] A. M. Lechner, G. M. Foody, and D. S. Boyd, "Applications in Remote Sensing to Forest Ecology and Management", *One Earth*, 2(5), 2020, 405–12.

[19] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention Network for Semantic Segmentation of Fine Resolution Remote Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2021, 1–13.

[20] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery", *ISPRS journal of photogrammetry and remote sensing*, 181, 2021, 84–98.

[21] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual State Space Model", *arXiv preprint arXiv:2401.10166*, 2024.

[22] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking Transformers for Semantic Segmentation of Remote Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, 61, 2023, 1–15, DOI: 10.1109/TGRS.2023.3302024.

[23] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "SAM-Assisted Remote Sensing Imagery Semantic Segmentation With Object and Boundary Constraints", *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024, 1–16, DOI: 10.1109/TGRS.2024.3443420.

[24] X. Ma, X. Zhang, and M.-O. Pun, "A Crossmodal Multiscale Fusion Network for Semantic Segmentation of Remote Sensing Data", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 2022, 3463–74.

[25] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation", *IEEE Geoscience and Remote Sensing Letters*, 21, 2024, 1–5, DOI: 10.1109/LGRS.2024.3414293.

[26] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation", *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024, 1–15, DOI: 10.1109/TGRS.2024.3373033.

[27] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised Domain Adaptation Augmented by Mutually Boosted Attention for Semantic Segmentation of VHR Remote Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, 61, 2023, 1–15.

[28] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review", *ISPRS journal of photogrammetry and remote sensing*, 66(3), 2011, 247–59.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 779–88.

[30] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", in, Computational and Biological Learning Society, 2015, 1–14.

[31] A. H. Strahler, "The use of prior probabilities in maximum likelihood classification of remotely sensed data", *Remote sensing of Environment*, 10(2), 1980, 135–63.

[32] J. Sui, X. Ma, X. Zhang, and M.-O. Pun, "GCRDN: Global Context-Driven Residual Dense Network for Remote Sensing Image Superresolution", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 2023, 4457–68, DOI: 10.1109/JSTARS.2023.3273081.

[33] J. Sui, Y. Ma, W. Yang, X. Zhang, M.-O. Pun, and J. Liu, "Diffusion Enhancement for Cloud Removal in Ultra-Resolution Remote Sensing Imagery", *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024, 1–14, DOI: 10.1109/TGRS.2024.3411671.

[34] H. Sun, J. Liu, J. Yang, and Z. Wu, "HMAFNet: Hybrid Mamba-Attention Fusion Network for Remote Sensing Image Semantic Segmentation", *IEEE Geoscience and Remote Sensing Letters*, 22, 2025, 1–5, DOI: 10.1109/LGRS.2025.3554786.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,. Kaiser, and I. Polosukhin, "Attention Is All You Need", *Advances in neural information processing systems*, 30, 2017.

[36] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, *LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation*, Zenodo, October 2021, DOI: 10.5281/zenodo.5706578, https://doi.org/10.5281/zenodo.5706578.

[37] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A Re-
      mote Sensing Land-Cover Dataset for Domain Adaptive Semantic Seg-
      mentation", in *Proceedings of the Neural Information Processing Sys-
      tems Track on Datasets and Benchmarks*, ed. J. Vanschoren and S.
      Yeung, Vol. 1, Curran Associates, Inc., 2021, https : / / datasets -
      benchmarks - proceedings . neurips . cc / paper _ files / paper / 2021 / file /
      4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf.

[38] L. Wang, D. Li, S. Dong, X. Meng, X. Zhang, and D. Hong, "Pyramid-
      Mamba: Rethinking Pyramid Feature Fusion with Selective Space State
      Model for Semantic Segmentation of Remote Sensing Imagery", *arXiv
      preprint arXiv:2406.10828*, 2024.

[39] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkin-
      son, "UNetFormer: A UNet-like Transformer for Efficient Semantic Seg-
      mentation of Remote Sensing Urban Scene Imagery", *ISPRS Journal of
      Photogrammetry and Remote Sensing*, 190, 2022, 196–214.

[40] M. Wójtowicz, A. Wójtowicz, J. Piekarczyk, *et al.*, "Application of re-
      mote sensing methods in agriculture", *Communications in biometry and
      crop science*, 11(1), 2016, 31–50.

[41] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN
      and Multiscale Transformer Fusion Network for Remote-Sensing Image
      Semantic Segmentation", *IEEE Transactions on Geoscience and Remote
      Sensing*, 61, 2023, 1–12, DOI: 10.1109/TGRS.2023.3314641.

[42] S. Zhao, Q. Wang, Y. Li, S. Liu, Z. Wang, L. Zhu, and Z. Wang, "An
      overview of satellite remote sensing technology used in Chinas environ-
      mental protection", *Earth Science Informatics*, 10, 2017, 137–48.

[43] S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, and W. Ouyang, "RS-
      Mamba for Large Remote Sensing Image Dense Prediction", *IEEE
      Transactions on Geoscience and Remote Sensing*, 62, 2024, 1–14, DOI:
      10.1109/TGRS.2024.3425540.

[44] H. Zhou, X. Xiao, H. Li, X. Liu, and P. Liang, "Hybrid Shunted Trans-
      former embedding UNet for remote sensing image semantic segmenta-
      tion", *Neural Computing and Applications*, 2024, 1–16.

[45] W. Zhou, S.-I. Kamata, H. Wang, M.-S. Wong, *et al.*, "Mamba-
      in-Mamba: Centralized Mamba-Cross-Scan in Tokenized M amba
      Model for Hyperspectral Image Classification", *arXiv preprint
      arXiv:2405.12003*, 2024.

[46] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision
      Mamba: Efficient Visual Representation Learning with Bidirectional
      State Space Model", *arXiv preprint arXiv:2401.09417*, 2024.

[47] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraun-
      dorfer, "Deep Learning in Remote Sensing: A Comprehensive Review
      and List of Resources", *IEEE geoscience and remote sensing magazine*,
      5(4), 2017, 8–36.