

Article

CoMS-UNet: A Cohesive Multi-Scale Network for Semantic Reorganization and Scale-Aware Context Modeling in Remote Sensing Image Segmentation

Yankai Wang , Shaochen Jiang ^{*} , Liejun Wang ^{*}  and Beibei Gao 

School of Computer Science and Technology, Xinjiang University, Huarui Street, Urumqi 830046, China; 107552503687@stu.xju.edu.cn (Y.W.); 107552403898@stu.xju.edu.cn (B.G.)

^{*} Correspondence: scj@xju.edu.cn (S.J.); wljxju@xju.edu.cn (L.W.)

Abstract

Semantic segmentation is essential for fine-grained land-cover interpretation from high-resolution remote sensing imagery. However, existing methods still show limitations in coordinating multi-stage encoder features, providing scale-aware contextual guidance at the bottleneck layer, and facilitating cross-channel information exchange during decoder reconstruction. To address these issues, this paper proposes CoMS-UNet, a cohesive multi-scale U-Net that integrates semantic reorganization, scale-aware context modeling, and channel-aware decoder reconstruction. In encoder–decoder skip connections, the Multi-Scale Semantic Reorganization Module (MSRM) reorganizes multi-stage encoder features through a fusion–separation–refusion process to achieve cross-scale semantic alignment. At the bottleneck stage, the Scale-Aware Module (SAM) constructs predefined receptive-field branches and adaptively weights their contextual responses to enhance scale-relevant semantic guidance. During decoder reconstruction, the Channel-Spatial Shuffle Mamba Block (CSSMBlock) promotes cross-channel information exchange and long-range spatial dependency modeling to improve structural detail recovery. Through this unified feature-processing framework, CoMS-UNet strengthens semantic consistency, scale-aware context aggregation, and channel-aware decoder reconstruction. Experiments on the ISPRS Vaihingen dataset and the labeled LoveDA validation set show that CoMS-UNet achieves a mean mIoU of $86.11 \pm 0.35\%$ over three independent runs on Vaihingen and an mIoU of 52.48% on LoveDA.

Keywords: remote sensing semantic segmentation; high-resolution remote sensing imagery; multi-scale semantic reorganization; scale-aware learning; land-cover interpretation



Academic Editor: Jiachen Yang

Received: 16 June 2026

Revised: 30 July 2026

Accepted: 4 August 2026

Published: 6 August 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Remote sensing semantic segmentation aims to assign a semantic category to each pixel in a remote sensing image, thereby enabling detailed interpretation of land cover and target distribution. With the increasing availability of high-resolution Earth observation imagery, remote sensing semantic segmentation has become increasingly important in crop classification [1], agricultural parcel delineation [2], coastal wetland monitoring [3], and post-disaster damage assessment [4]. Compared with conventional computer vision datasets, high-resolution remote sensing datasets generally contain more complex backgrounds and greater object-scale variation. Existing semantic segmentation methods, especially convolutional encoder–decoder architectures, have improved pixel-wise segmentation accuracy by extracting hierarchical features and using decoder structures to

recover spatial details. However, remote sensing datasets typically exhibit complex spatial layouts, significant object-scale variations, spectral differences, and blurred boundaries. These factors may lead to fragmented land-cover boundaries, confusion between visually similar categories, and incomplete extraction of small or narrow objects in fine-grained land-cover interpretation. Therefore, improving the model's ability to maintain semantic consistency, adapt to scale variations, and recover fine structural details remains essential for reliable interpretation of high-resolution remote sensing imagery.

Convolutional Neural Networks (CNNs) [5,6] have laid an important foundation for semantic segmentation because they can extract hierarchical visual representations from images. On this basis, Fully Convolutional Networks (FCNs) [7] perform dense pixel-wise prediction by replacing the fully connected layers with convolutional operations, while U-Net [8] builds an encoder–decoder architecture and uses skip connections to recover spatial details, making it widely adopted in segmentation tasks. Attention U-Net [9] further enhances the skip pathways by introducing attention gates, which filter encoder features before their fusion with decoder features. In addition to these encoder–decoder designs, multi-scale contextual modeling has been widely studied. DeepLabV3+ [10] relies on Atrous Spatial Pyramid Pooling (ASPP), whereas PSPNet [11] adopts Pyramid Pooling Module (PPM), allowing high-level contextual features to be captured from different receptive-field ranges. Depthwise Separable Convolution (DSC) improves computational efficiency by decoupling spatial filtering and channel mixing, and ShuffleNet [12] uses channel shuffle to promote information exchange among channels. With the emergence of state-space models and the Mamba architecture, CM-UNet [13] integrates CNN-based local feature extraction, Mamba-based long-range dependency modeling, and a Multi-Scale Attention Aggregation (MSAA) module to fuse multi-scale encoder features for remote sensing segmentation.

Against this background, recent studies have further explored boundary-aware multi-scale representation and state-space modeling for remote sensing image segmentation. BEMS-UNetFormer [14] emphasizes boundary enhancement and multi-scale context aggregation, whereas MLMFPN [15] focuses on local–global synergy and cross-scale context interaction. CVMH-UNet [16], PyramidMamba [17], and SwinMamba [18] further combine state-space modeling with multi-scale multi-frequency fusion, pyramid feature fusion, and hybrid local–global modeling, respectively. These advances highlight the continuing importance of multi-scale representation, local detail modeling, and long-range dependency modeling in high-resolution remote sensing image segmentation.

Despite these advances, several challenges remain in fine-grained land-cover interpretation from high-resolution remote sensing imagery: (1) Multi-stage encoder features differ in spatial detail and semantic abstraction, and insufficient coordination may cause fragmented predictions. (2) Fixed bottleneck aggregation may not adequately adapt to substantial object-scale variation. (3) Limited inter-channel interaction during decoding may weaken boundary and structural detail reconstruction.

The main contributions of this paper are summarized as follows:

- We propose the Multi-Scale Semantic Reorganization Module (MSRM), which strengthens cross-scale semantic connections and promotes semantic reorganization of multi-stage encoder features in skip connections. By adopting a fusion–separation–refusion strategy, MSRM aligns, fuses, and reassembles hierarchical encoder features, thereby alleviating the interference caused by noisy shallow spatial details and the spatial detail degradation of deep contextual representations in high-resolution remote sensing scenes.
- We design a Scale-Aware Module (SAM) to improve the model's sensitivity to target scale variations in remote sensing imagery. SAM constructs receptive-field branches with predefined dilation rates and adaptively weights their contextual responses according to the input feature. By learning input-dependent branch weights at the

bottleneck layer, SAM selectively emphasizes scale-relevant contextual cues, thereby providing semantic guidance for decoder reconstruction.

- We propose the Channel-Spatial Shuffle Mamba Block (CSSMBlock) to strengthen channel-aware decoder reconstruction, thereby facilitating the effective recovery of fine structural details in remote sensing imagery. CSSMBlock integrates channel shuffle-based local extraction, SS2D-based long-range modeling, and dual attention to alleviate the channel-wise information isolation caused by depthwise separable convolution. By promoting cross-channel joint modeling during decoder reconstruction, CSSMBlock improves the recovery of object boundaries, narrow structures, and fragmented land-cover regions.

2. Materials and Methods

2.1. Architecture Overview

The proposed CoMS-UNet (Figure 1) follows an encoder–bottleneck–decoder structure and forms a cohesive feature-processing framework through multi-scale semantic alignment, adaptive bottleneck aggregation, and channel-aware decoder reconstruction. To improve the semantic coordination of multi-stage encoder features in skip connections, MSRM introduces learnable cross-level interaction instead of direct feature stacking. This design promotes the complementary integration of shallow spatial details and deep contextual representations, thereby strengthening cross-scale semantic connections and improving the semantic organization of multi-scale features. We use the Enhanced Channel-Shuffle Block (ECSBlock) and the Local Channel-Shuffle Block (LCSBlock) in SAM and CSSMBlock, respectively. The specific structures of ECSBlock and LCSBlock are shown in Figure 2a and Figure 2b, respectively. SAM addresses the fixed aggregation pattern at the bottleneck stage by constructing ECSBlock-based receptive-field branches with predefined dilation rates and adaptively weighting their contextual responses according to the input feature. As for the decoder, our proposed CSSMBlock internally replaces the standard DSC with LCSBlock. Through the channel shuffle operation, LCSBlock facilitates cross-channel information exchange, alleviates the channel-wise information isolation caused by depthwise convolution, and helps maintain feature coherence during decoder reconstruction.

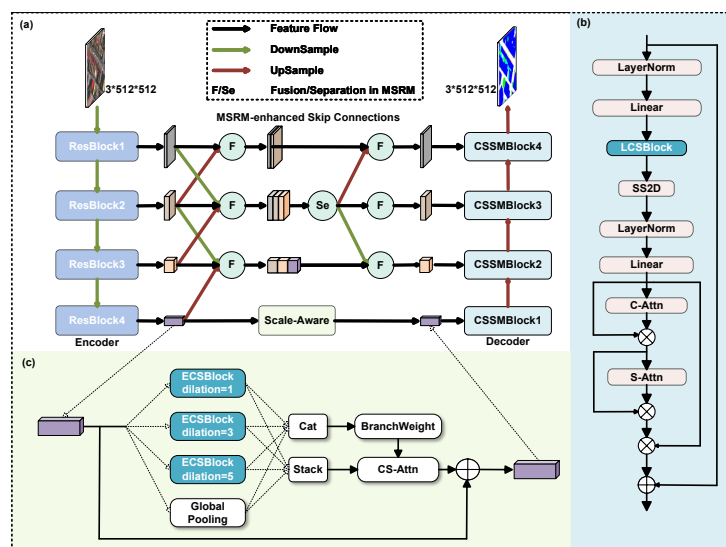


Figure 1. Overall architecture of CoMS-UNet. (a) CoMS-UNet. (b) CSSMBlock. (c) Scale-Aware Module. F and Se indicate the fusion and separation operations in MSRM, respectively. C-Attn, S-Attn, and CS-Attn denote channel, spatial, and channel-spatial attention, respectively. The Enhanced Channel-Shuffle Block (ECSBlock) is used to construct multi-scale receptive-field branches in SAM, while the Local Channel-Shuffle Block (LCSBlock) is used for local feature extraction in CSSMBlock.

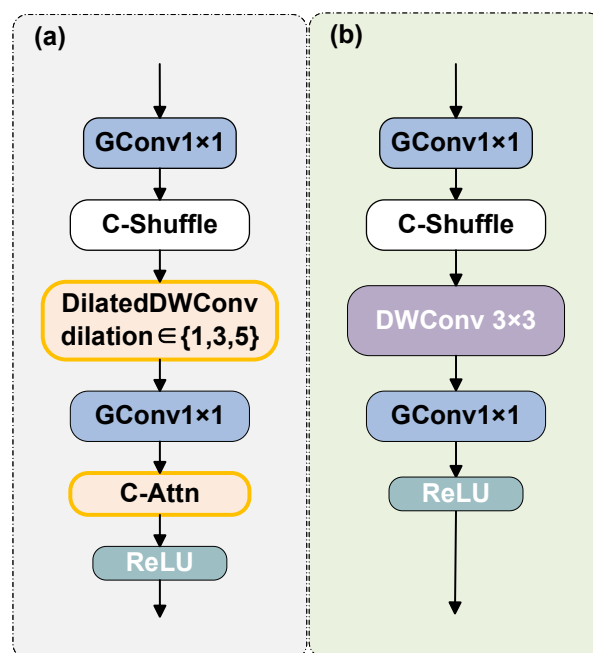


Figure 2. Architectural details of the proposed feature blocks. (a) Enhanced Channel-Shuffle Block (ECSBlock). (b) Local Channel-Shuffle Block (LCSBlock). Three ECSBlock instances with dilation rates of 1, 3, and 5 are used to construct the receptive-field branches in SAM. LCSBlock is used in the decoder and adopts a standard 3×3 depthwise convolution. GConv1 $\times$ 1, C-Shuffle, DWConv, and C-Attn denote grouped 1×1 convolution, channel shuffle, depthwise convolution, and channel attention, respectively.

2.2. Multi-Scale Semantic Reorganization Module (MSRM)

CoMS-UNet introduces a Multi-Scale Semantic Reorganization Module (MSRM) to improve cross-scale semantic coordination in skip connections. MSRM reorganizes hierarchical encoder features through a fusion-separation-refusion process. This design strengthens cross-scale semantic connections and enables more coherent integration of shallow spatial details and deep contextual representations.

Given the hierarchical encoder features $\{E_i\}_{i=1}^4$, MSRM first aligns neighboring features to the target scale using 1×1 channel projection and bilinear interpolation. Let $\tilde{E}_{j \rightarrow i}$ denote the feature from level j after being aligned to level i . The preliminary fusion stage aggregates complementary contextual information from adjacent levels through channel-wise concatenation:

$$F_1 = \mathcal{F}_1([E_1, \tilde{E}_{2 \rightarrow 1}]), \quad (1)$$

$$F_2 = \mathcal{F}_2([\tilde{E}_{1 \rightarrow 2}, E_2, \tilde{E}_{3 \rightarrow 2}]), \quad (2)$$

$$F_3 = \mathcal{F}_3([\tilde{E}_{2 \rightarrow 3}, E_3, \tilde{E}_{4 \rightarrow 3}]), \quad (3)$$

The fusion function $\mathcal{F}_i$ consists of two stacked 3×3 convolutions, batch normalization, ReLU activation, and a channel attention mechanism that dynamically recalibrates channel weights via global average pooling and a two-layer fully connected bottleneck.

To obtain a balanced basis for subsequent feature separation, F_2 is used as the central hub between semantic abstraction and spatial detail preservation. From this unified central hub, we apply three parallel separation branches to decouple scale-specific feature representations:

$$\mathbf{S}_1 = \mathcal{S}_1(\mathbf{F}_2), \quad (4)$$

$$\mathbf{S}_2 = \mathcal{S}_2(\mathbf{F}_2), \quad (5)$$

$$\mathbf{S}_3 = \mathcal{S}_3(\mathbf{F}_2), \quad (6)$$

Each of the separation functions $\mathcal{S}_i$ is implemented through a 3×3 convolution, batch normalization, ReLU, and the aforementioned channel attention mechanism. After these operations, we perform bilinear interpolation to resize the feature map to match the corresponding target resolution.

In the final fusion stage after separation, these separated features are recombined with the corresponding scaled features from the initial fusion. This integration incorporates multi-scale contextual information, enabling each level of feature to simultaneously possess high-level semantic information, while also retaining the spatial information of the shallow features:

$$\hat{\mathbf{E}}_1 = \mathcal{R}_1([\mathbf{F}_1, \tilde{\mathbf{S}}_1]), \quad (7)$$

$$\hat{\mathbf{E}}_2 = \mathbf{S}_2, \quad (8)$$

$$\hat{\mathbf{E}}_3 = \mathcal{R}_3([\mathbf{F}_3, \tilde{\mathbf{S}}_3]), \quad (9)$$

$\tilde{\mathbf{S}}_i$ represents the separated features after being aligned to the corresponding scale. The refinement function $\mathcal{R}_i$ has the same configuration structure as $\mathcal{F}_i$. Then, the obtained $\{\hat{\mathbf{E}}_1, \hat{\mathbf{E}}_2, \hat{\mathbf{E}}_3, \mathbf{E}_4\}$ is forwarded to the decoder.

2.3. Scale-Aware Module (SAM)

The Scale-Aware Module (SAM) is introduced at the bottleneck stage to enhance the scale-aware representation of deep features and provide more adaptive contextual guidance for decoder reconstruction. In high-resolution remote sensing scenes, fixed aggregation patterns may limit the ability of bottleneck features to emphasize scale-relevant contextual cues. To address this issue, SAM constructs multiple ECSBlock-based receptive-field branches and adaptively weights their contextual responses according to the input feature. Specifically, ECSBlock consists of a grouped 1×1 convolution, a channel shuffle operation for inter-group information exchange, a depthwise dilated convolution ($\text{DWConv}_{3 \times 3}^d$) for spatial context modeling, a final grouped 1×1 convolution, and a channel attention mechanism for channel-wise recalibration. This structure refines each branch feature before multi-branch aggregation.

Based on this structure, SAM adopts four parallel paths. The first three paths employ ECSBlocks with predefined dilation rates $d \in \{1, 3, 5\}$, while the fourth path introduces image-level contextual information through global average pooling (GAP). Let $\mathcal{H}_d$ denote the ECSBlock with dilation rate d , and let $\mathbf{X}$ denote the bottleneck input feature:

$$\mathbf{B}_1 = \mathcal{H}_1(\mathbf{X}), \quad \mathbf{B}_2 = \mathcal{H}_3(\mathbf{X}), \quad \mathbf{B}_3 = \mathcal{H}_5(\mathbf{X}), \quad (10)$$

With a 3×3 depthwise dilated convolution, the three branches correspond to effective receptive fields of 3×3 , 7×7 , and 11×11 , respectively. These branches allow the bottleneck feature to capture contextual responses from different spatial ranges, including fine local details, medium-range patterns, and larger-range contextual cues. To further complement these receptive-field branches, the fourth branch introduces image-level contextual information through global average pooling (GAP):

$$\mathbf{B}_4 = \mathcal{U}(\delta(\text{BN}(\text{Conv}_{1 \times 1}(\text{GAP}(\mathbf{X}))))), \quad (11)$$

where $\mathcal{U}$ denotes bilinear upsampling, BN is batch normalization, and δ is the ReLU activation.

Rather than simply concatenating or summing the branch outputs, SAM explicitly models the relative importance of different scale responses through an input-dependent branch weighting mechanism. The concatenated multi-branch feature is denoted as $\mathbf{B}_{\text{cat}} = [\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \mathbf{B}_4]$. It is then used to generate adaptive branch weights $\alpha \in \mathbb{R}^4$:

$$\alpha = \text{Softmax}(\text{Conv}_{1 \times 1}(\text{GAP}(\delta(\text{BN}(\text{Conv}_{1 \times 1}(\mathbf{B}_{\text{cat}})))))). \quad (12)$$

The branch-weighted feature is obtained by aggregating the contextual responses from the four paths:

$$\mathbf{F} = \sum_{i=1}^4 \alpha_i \mathbf{B}_i. \quad (13)$$

The aggregated feature is further refined by a joint channel-spatial attention module, denoted as $\mathcal{A}_{\text{cs}}$. Finally, the attended feature is projected by a 1×1 convolution and combined with the original bottleneck input through a residual connection:

$$\hat{\mathbf{O}} = \delta(\text{BN}(\text{Conv}_{1 \times 1}(\mathcal{A}_{\text{cs}}(\mathbf{F})))) + \mathbf{X}. \quad (14)$$

Through predefined receptive-field branches, input-dependent branch weighting, channel-spatial refinement, and residual learning, SAM enables bottleneck features to emphasize scale-relevant contextual cues. This process improves the scale-aware semantic guidance delivered from the bottleneck layer to the decoder.

2.4. Channel-Spatial Shuffle Mamba Block (CSSMBlock)

CSSMBlock is designed to enhance channel-aware decoder reconstruction by promoting inter-channel information exchange and structural detail recovery. To this end, it follows a sequential refinement process consisting of LCSBlock-based local modeling, SS2D-based long-range dependency modeling, and channel-spatial attention refinement.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, CSSMBlock first reshapes it into a sequence representation $\mathbf{X}_{\text{seq}}$. Then, layer normalization and linear projection are applied to expand the channel dimension:

$$\mathbf{Z}_1 = \text{Linear}_{\text{up}}(\text{LN}(\mathbf{X}_{\text{seq}})), \quad (15)$$

where $\mathbf{Z}_1$ has an expanded channel dimension with an expansion factor $\lambda = 2$.

The expanded sequence is reshaped back into a spatial feature map and then processed by LCSBlock for local feature extraction, where $\mathcal{H}_{\text{LCS}}$ denotes the LCSBlock transformation:

$$\mathbf{Z}_2 = \mathcal{H}_{\text{LCS}}(\text{Reshape}(\mathbf{Z}_1)). \quad (16)$$

LCSBlock employs grouped 1×1 convolution and channel shuffle to facilitate information exchange across channel groups. Different from the ECSBlock used in the bottleneck stage, LCSBlock adopts a standard non-dilated 3×3 depthwise convolution and does not include an internal channel attention mechanism. This design focuses on local structural modeling while alleviating the channel-wise information isolation caused by the depthwise convolution in DSC.

To further model long-range spatial dependencies beyond local convolutional extraction, the locally encoded feature is processed by the two-dimensional state space model, denoted as SS2D:

$$\mathbf{Z}_3 = \mathcal{M}_{\text{SS2D}}(\mathbf{Z}_2). \quad (17)$$

Through two-dimensional selective scanning, SS2D models spatial dependencies across distant regions and enriches the structural context required for decoder reconstruction. The resulting feature is then normalized and projected back to the original channel dimension:

$$\mathbf{Z}_4 = \text{Linear}_{\text{down}}(\text{LN}(\mathbf{Z}_3)). \quad (18)$$

After the channel projection, $\mathbf{Z}_4$ is reshaped back to the spatial form for the subsequent channel-spatial attention refinement.

CSSMBlock further introduces a dual attention mechanism to refine decoder features from both channel and spatial dimensions. The attention-refined features are formulated as

$$\mathbf{Z}_c = \mathcal{A}_c(\mathbf{Z}_4) \odot \mathbf{Z}_4, \quad (19)$$

$$\mathbf{Z}_{cs} = \mathcal{A}_s(\mathbf{Z}_c) \odot \mathbf{Z}_c, \quad (20)$$

$$\mathbf{Z}_f = \mathbf{Z}_4 \odot \mathbf{Z}_{cs}, \quad (21)$$

where $\mathcal{A}_c$ and $\mathcal{A}_s$ denote the channel attention and spatial attention operations, respectively, $\mathbf{Z}_f$ denotes the fused attention-refined feature, and $\odot$ represents element-wise multiplication.

Finally, a residual connection is introduced to preserve the original feature information and stabilize decoder feature reconstruction:

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z}_f. \quad (22)$$

Through the progressive refinement from local structural modeling to long-range dependency modeling and channel-spatial attention, CSSMBlock strengthens cross-channel joint modeling during decoder reconstruction. This improves the recovery of object boundaries and fine structural details.

3. Results

3.1. Experimental Setup

3.1.1. Dataset Settings

ISPRS Vaihingen Dataset. The ISPRS Vaihingen benchmark includes 33 high-resolution aerial images acquired over Vaihingen, Germany, with a ground sampling distance of 9 cm. In our experiments, 16 images are assigned to the training set, and the other 17 images are used for testing.

LoveDA Dataset. LoveDA is a high-spatial-resolution remote sensing benchmark for land-cover semantic segmentation and unsupervised domain adaptation. We use the official training split, comprising 2522 images, for model training and the labeled validation split, comprising 1669 images, for evaluation. Since the labels of the official test set are not publicly available, all LoveDA results are reported on the labeled validation set.

3.1.2. Implementation Details

The experiments were implemented in PyTorch 2.0 and conducted on a single NVIDIA A40 GPU (NVIDIA Corporation, Santa Clara, CA, USA). For CoMS-UNet, the encoder was initialized with pretrained ResNet-18 weights. The model was trained for 105 epochs with a batch size of 8. AdamW combined with Lookahead was used for optimization. The learning rate was annealed from 6×10^{-4} to 1×10^{-5} using a cosine schedule, and the weight decay was set to 0.01. The training objective combined soft cross-entropy loss, Dice loss, and auxiliary supervision from three decoder stages. Each auxiliary loss was assigned a weight of 0.4.

Training samples were generated by random scaling followed by content-balanced cropping into 512-by-512-pixel patches. The maximum proportion of the dominant class

within a crop was set to 0.75. Scale factors of 0.5, 0.75, 1.0, 1.25, and 1.5 were used for ISPRS Vaihingen. For LoveDA, the corresponding factors were 0.75, 1.0, 1.25, and 1.5. Four-image mosaic augmentation was applied to both datasets with a probability of 0.25. Random 90-degree rotation was applied to ISPRS Vaihingen, whereas random horizontal flipping was applied to LoveDA. Each dataset-specific transformation was used with a probability of 0.5.

For CoMS-UNet, inference on the ISPRS Vaihingen dataset was performed using offline-generated 1024-by-1024-pixel tiles. LoveDA images were evaluated directly at their native 1024-by-1024-pixel resolution. Each input was processed as a whole, without additional online sliding-window partitioning. Multi-scale and flip-based test-time augmentation was applied, and the resulting prediction scores were averaged to obtain the final output.

3.1.3. Evaluation Metrics and Comparison Protocol

To compare CoMS-UNet with existing methods, the mean F1 score (mF1), mean Intersection over Union (mIoU), and overall accuracy (OA) are used as evaluation metrics. Among them, mF1 reflects the balance of precision and recall across different classes, mIoU more strictly measures the overlap between the predicted labels and the true annotations, and OA evaluates the pixel-wise classification accuracy of the model. By comprehensively considering these three metrics, the performance of CoMS-UNet can be evaluated from the category level and the global level. For the ISPRS Vaihingen dataset, the reported scores were computed using eroded reference boundaries, and boundary pixels were ignored. The clutter class was included in the confusion matrix and the calculation of overall accuracy (OA). Its individual IoU and F1 scores were computed but are not reported in Table 1, and it was excluded from the calculation of mIoU and mean F1.

All comparative methods reported in Tables 1 and 2 were reproduced in our experimental environment. Model-specific implementation settings, including training hyperparameters, input preprocessing, inference procedures, and test-time augmentation, were based on the corresponding publications or official implementations. Within each dataset, all methods used the same data split, evaluation images, reference labels, and metric implementation. The citations accompanying the method names refer to the corresponding original publications.

Table 1. Experimental results (%) on the ISPRS Vaihingen dataset. Class-wise values denote F1 scores, and bold values indicate the results of CoMS-UNet.

Method	Backbone	Imp.	Build.	Low.v	Tree	Car	mF1	OA	mIoU
ABCNet [19]	R18	96.25	94.41	83.96	89.94	81.12	89.14	92.56	80.90
BANet [20]	ResT	96.61	94.88	84.20	89.34	83.35	89.68	92.75	81.72
Segmenter [21]	ViT-T	95.33	93.61	76.31	88.05	58.87	82.43	91.45	72.22
UNetFormer [22]	R18	96.60	95.45	83.44	89.39	86.83	90.34	92.88	82.77
FTUNetFormer [22]	Swin-B	97.01	95.76	84.57	89.99	88.62	91.19	93.39	84.14
CMTFNet [23]	R50	96.13	94.43	83.66	89.20	85.62	89.81	92.33	81.86
RS3Mamba [24]	R18-Mamba	96.77	96.10	81.97	91.11	89.72	91.13	94.05	84.14
CM-UNet [13]	R18	97.14	96.51	82.67	91.73	91.00	91.81	94.49	85.28
CoMS-UNet (Ours)	R18	97.46	96.75	84.58	92.42	91.68	92.58	95.01	86.51

The main comparison results were obtained from runs using random seed 42. To examine run-to-run variation, each configuration in the ablation study was independently trained using random seeds 0, 1, and 42. Results from these runs are reported as the mean $\pm$ standard deviation.

Table 2. Experimental results (%) on the labeled LoveDA validation set. Class-wise values denote IoU scores, and bold values indicate the results of CoMS-UNet.

Method	Backbone	Bkg.	Build.	Road	Water	Barr.	Forest	Agri.	mIoU
UNetFormer [22]	R18	54.12	59.58	51.76	63.50	33.35	43.97	50.20	50.93
RS3Mamba [24]	R18-Mamba	50.67	59.29	51.97	52.71	28.26	42.27	50.69	47.98
Segmenter [21]	ViT-T	51.83	55.58	49.95	71.06	24.12	39.62	59.73	50.27
ABCNet [19]	R18	51.97	59.26	52.54	62.97	30.39	37.22	45.82	48.59
CM-UNet [13]	R18	53.31	62.86	53.42	64.28	34.40	41.48	50.51	51.47
CoMS-UNet (Ours)	R18	54.21	63.79	53.16	65.85	34.77	44.20	51.35	52.48

3.2. Results on ISPRS Vaihingen Dataset

Table 1 presents the quantitative results on the ISPRS Vaihingen dataset. CoMS-UNet obtains an mF1 of 92.58%, an OA of 95.01%, and an mIoU of 86.51%. Compared with the CM-UNet baseline, CoMS-UNet improves mF1, OA, and mIoU by 0.77, 0.52, and 1.23 percentage points, respectively. Improvements are also observed in the F1 scores of all five land-cover classes.

Figure 3 presents a visual comparison between CM-UNet and CoMS-UNet on the ISPRS Vaihingen dataset. Compared with CM-UNet, CoMS-UNet produces more continuous segmentation regions and preserves object boundaries and small targets more completely in several examples. Among the class-wise results, the F1 scores for Low vegetation, Tree, and Car increase by 1.91, 0.69, and 0.68 percentage points, respectively. These results indicate that CoMS-UNet reduces some fragmented predictions and improves the segmentation of vegetation regions and small objects.

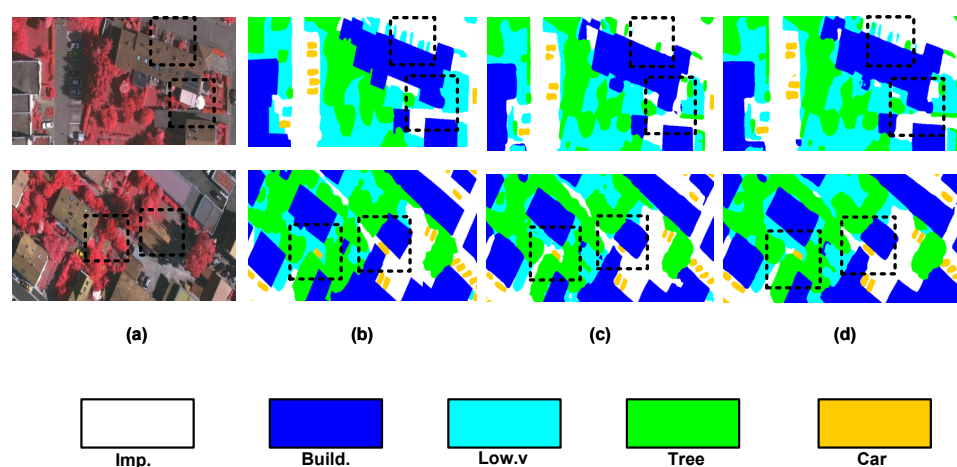


Figure 3. Visual comparisons on the ISPRS Vaihingen dataset. Columns (a–d) show RGB images, ground truth, CM-UNet, and CoMS-UNet (Ours), respectively.

3.3. Results on LoveDA Dataset

Table 2 presents the quantitative results on the labeled LoveDA validation set. CoMS-UNet obtains an mIoU of 52.48%, which is 1.01 percentage points higher than that of the CM-UNet baseline. The results indicate that the modifications introduced in CoMS-UNet provide an improvement over the baseline on the LoveDA dataset.

Figure 4 presents a visual comparison between CM-UNet and CoMS-UNet on the labeled LoveDA validation set. Compared with CM-UNet, CoMS-UNet produces more coherent segmentation regions in several examples and better preserves some land-cover boundaries. Among the class-wise results, the IoU scores for Forest, Water, Building, Background, and Agriculture increase by 2.72, 1.57, 0.93, 0.90, and 0.84 percentage points,

respectively. In contrast, the IoU for Road decreases slightly by 0.26 percentage points. These results indicate that CoMS-UNet improves the segmentation of several land-cover categories, although the improvement is not uniform across all classes.

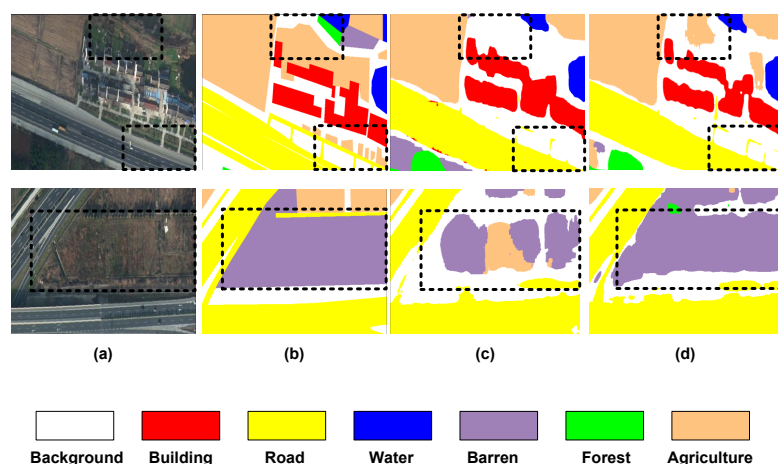


Figure 4. Results on the labeled LoveDA validation set. Columns (a–d) show RGB images, ground truth, CM-UNet, and CoMS-UNet (Ours), respectively.

3.4. Ablation Study

Table 3 presents the ablation results on the ISPRS Vaihingen dataset. Compared with the baseline mIoU of $84.88 \pm 0.41\%$, the MSRM-only, SAM-only, and CSSMBlock-only variants achieve mean mIoU scores of $85.93 \pm 0.19\%$, $85.49 \pm 0.60\%$, and $85.26 \pm 0.40\%$, respectively. Among the single-module configurations, MSRM provides the largest mean mIoU increase. The complete CoMS-UNet achieves the highest mean mF1 and mIoU, reaching $92.33 \pm 0.22\%$ and $86.11 \pm 0.35\%$, respectively, with a mean mIoU improvement of 1.23 percentage points over the baseline. Although the MSRM-only variant obtains a slightly higher mean OA than the complete model, the difference is only 0.03 percentage points and is small relative to the observed run-to-run variation. Therefore, the two configurations show similar OA results.

Among the two-module configurations, MSRM+SAM achieves the highest mean mIoU of $86.05 \pm 0.12\%$, followed by MSRM+CSSMBlock with $85.84 \pm 0.02\%$ and SAM+CSSMBlock with $85.48 \pm 0.31\%$. The complete CoMS-UNet obtains a mean mIoU of $86.11 \pm 0.35\%$. These results show that combining the modules does not produce strictly additive improvements, and the performance differences among several module combinations are relatively small.

To further demonstrate the effects of the proposed modules, Figure 5 presents qualitative ablation results on representative challenging regions from the ISPRS Vaihingen dataset. In the first row, the baseline produces fragmented predictions and incoherent object structures around buildings, whereas MSRM helps improve the semantic consistency of multi-stage encoder features and generate more coherent land-cover structures. In the second row, SAM recovers regions with different spatial sizes more completely, suggesting that it helps select more appropriate receptive-field responses. In the third row, CSSMBlock improves the reconstruction of visually similar structures, indicating that enhanced inter-channel interaction contributes to more consistent decoder reconstruction. These visual results provide qualitative evidence consistent with the quantitative ablation results in Table 3 and illustrate how the proposed modules alleviate different limitations in semantic coordination, scale-aware context modeling, and decoder reconstruction.

Table 3. Ablation study (%) on the ISPRS Vaihingen dataset. Results are reported as the mean $\pm$ standard deviation over three independent runs using random seeds 0, 1, and 42. Bold values indicate the best result for each evaluation metric.

MSRM	SAM	CSSMBlock	mF1	OA	mIoU
×	×	×	91.58 $\pm$ 0.24	94.50 $\pm$ 0.16	84.88 $\pm$ 0.41
✓	×	×	92.21 $\pm$ 0.10	94.86 $\pm$ 0.14	85.93 $\pm$ 0.19
×	✓	×	91.95 $\pm$ 0.37	94.69 $\pm$ 0.20	85.49 $\pm$ 0.60
×	×	✓	91.81 $\pm$ 0.24	94.62 $\pm$ 0.13	85.26 $\pm$ 0.40
✓	✓	×	92.29 $\pm$ 0.07	94.82 $\pm$ 0.08	86.05 $\pm$ 0.12
✓	×	✓	92.15 $\pm$ 0.02	94.76 $\pm$ 0.01	85.84 $\pm$ 0.02
×	✓	✓	91.93 $\pm$ 0.19	94.65 $\pm$ 0.06	85.48 $\pm$ 0.31
✓	✓	✓	92.33 $\pm$ 0.22	94.83 $\pm$ 0.16	86.11 $\pm$ 0.35

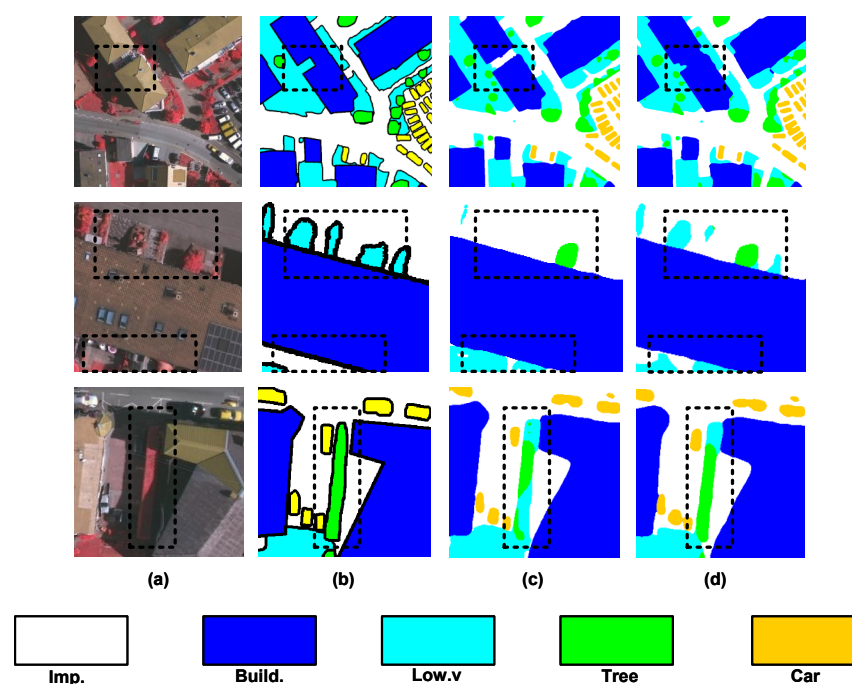


Figure 5. Qualitative ablation results on the ISPRS Vaihingen dataset. The three rows show the visualization results of the MSR-only, SAM-only, and CSSMBlock-only variants, respectively. Columns (a–d) show the RGB image, ground truth, CM-UNet baseline, and the corresponding single-module result, respectively.

3.5. Complexity and Efficiency Analysis

Table 4 compares the computational cost, parameter count, and segmentation accuracy of the evaluated models on the ISPRS Vaihingen dataset. Compared with CM-UNet, the FLOPs of CoMS-UNet increase from 12.02 to 38.48 GFLOPs at 512×512 and from 48.08 to 153.91 GFLOPs at 1024×1024 , while the parameter count increases from 12.89 M to 30.86 M. The additional computational cost is associated with the feature reorganization, multi-branch aggregation, and decoder reconstruction operations introduced in CoMS-UNet.

Compared with CM-UNet, CoMS-UNet increases the mIoU from 85.28% to 86.51%, corresponding to an improvement of 1.23 percentage points. However, this improvement is accompanied by approximately $3.2\times$ FLOPs and $2.39\times$ parameters, indicating a substantial increase in computational cost. Compared with CMTFNet, CoMS-UNet has a similar parameter count but requires more FLOPs and obtains a 4.65-percentage-point higher mIoU. Compared with FTUNetFormer, CoMS-UNet uses approximately one-third of its parameters and FLOPs while obtaining a 2.37-percentage-point higher mIoU. Overall, the current architecture places greater emphasis on segmentation accuracy than computational

efficiency, and reducing its computational cost and parameter count remains an important direction for future work.

Table 4. Computational complexity and parameter analysis. The mIoU values correspond to the runs using random seed 42, whereas FLOPs and parameter counts are independent of the random seed.

Model	FLOPs@512 ↓	FLOPs@1024 ↓	Param. ↓	mIoU (%) ↑
ABCNet [19]	15.63	62.50	13.39	80.90
UNetFormer [22]	11.74	46.97	11.69	82.77
FTUNetFormer [22]	126.30	499.19	96.14	84.14
CMTFNet [23]	33.07	131.87	30.07	81.86
CM-UNet [13]	12.02	48.08	12.89	85.28
CoMS-UNet (Ours)	38.48	153.91	30.86	86.51

FLOPs, Param., and mIoU denote floating-point operations, parameter count, and mean Intersection over Union, respectively. FLOPs are reported for a single RGB image at two resolutions: $1 \times 3 \times 512 \times 512$, corresponding to the training crop size, and $1 \times 3 \times 1024 \times 1024$, corresponding to the full-resolution inference tile. The reported FLOPs measure a single forward pass without test-time augmentation. The mIoU values are evaluated on the ISPRS Vaihingen dataset. Bold values indicate the results of CoMS-UNet.

4. Discussion

The results across the two datasets suggest that the modifications introduced at the skip connections, bottleneck, and decoder contribute to the segmentation performance of CoMS-UNet. However, the small difference between the complete model and the MSRM+SAM variant indicates that the effects of the modules are not strictly additive. Meanwhile, CoMS-UNet does not introduce a new segmentation paradigm, but modifies different stages of the feature-processing pipeline through semantic reorganization, scale-aware contextual modeling, and channel-aware reconstruction.

CoMS-UNet also has several limitations. Compared with CM-UNet, its FLOPs increase by approximately $3.2\times$ at both input resolutions, while the parameter count increases from 12.89 M to 30.86 M. Therefore, the accuracy improvement is accompanied by a substantial increase in computational cost. Performance also remains uneven across land-cover categories, particularly for narrow and continuous structures. Finally, the experiments are limited to two public datasets and do not fully establish generalization across different sensors, regions, seasons, and spatial resolutions. Future work will therefore focus on reducing model complexity, improving the segmentation of narrow structures and small objects, and evaluating the model under broader cross-domain and multimodal settings.

5. Conclusions

In this study, we present CoMS-UNet for remote sensing image semantic segmentation by integrating multi-scale semantic reorganization, scale-aware contextual modeling, and channel-aware decoder reconstruction. MSRM reorganizes multi-stage encoder features, SAM provides scale-aware contextual information at the bottleneck layer, and CSSMBlock enhances inter-channel interaction and spatial dependency modeling during decoder reconstruction. On the ISPRS Vaihingen dataset, CoMS-UNet achieves a mean mIoU of $86.11 \pm 0.35\%$ over three independent runs, representing a mean improvement of 1.23 percentage points over the CM-UNet baseline. On the labeled LoveDA validation set, CoMS-UNet obtains an mIoU of 52.48%. These results indicate that the proposed feature-processing design across the skip connections, bottleneck, and decoder can improve segmentation accuracy. However, the improvement is accompanied by increased computational cost and parameter count. Future work will prioritize model simplification and broaden the scope of applications in remote sensing.

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

Funding: This research was funded by the Data Element-Based Information System for Unmanned Platforms for Low-Altitude Inspection project, grant number 2025YFF0515604.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ISPRS Vaihingen dataset can be accessed via the official ISPRS benchmark repository (<https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/>, accessed on 15 June 2026). The LoveDA dataset is available in the Zenodo repository (<https://zenodo.org/record/5706578>, accessed on 15 June 2026).

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

Abbreviations

The following abbreviations are used in this manuscript:

CoMS-UNet	Cohesive Multi-Scale U-Net
MSRM	Multi-Scale Semantic Reorganization Module
SAM	Scale-Aware Module
CSSMBlock	Channel-Spatial Shuffle Mamba Block
ECSBlock	Enhanced Channel-Shuffle Block
LCSBlock	Local Channel-Shuffle Block
DSC	Depthwise Separable Convolution
SS2D	Two-Dimensional Selective Scan
MSAA	Multi-Scale Attention Aggregation
GAP	Global Average Pooling
BN	Batch Normalization
LN	Layer Normalization
GConv	Grouped Convolution
DWConv	Depthwise Convolution
C-Shuffle	Channel Shuffle
C-Attn	Channel Attention
S-Attn	Spatial Attention
CS-Attn	Channel-Spatial Attention
FLOPs	Floating-Point Operations
RGB	Red–Green–Blue

References

1. Zhang, C.; Kerner, H.; Wang, S.; Hao, P.; Li, Z.; Hunt, K.A.; Abernethy, J.; Zhao, H.; Gao, F.; Di, L.; et al. Remote sensing for crop mapping: A perspective on current and future crop-specific land cover data products. *Remote. Sens. Environ.* **2025**, *330*, 114995. [CrossRef]
2. Zheng, J.; Ye, Z.; Wen, Y.; Huang, J.; Zhang, Z.; Li, Q.; Hu, Q.; Xu, B.; Zhao, L.; Fu, H. A comprehensive review of agricultural parcel and boundary delineation from remote sensing images: Recent progress and future perspectives. *IEEE Geosci. Remote. Sens. Mag.* **2026**, *14*, 206–237. [CrossRef]
3. Nuradili, P.; Zhou, J.; Zhou, G.; Melgani, F. Deep learning method for wetland segmentation in unmanned aerial vehicle multispectral imagery. *Remote. Sens.* **2024**, *16*, 4777. [CrossRef]
4. Hertel, V.; Geiß, C.; Wieland, M.; Taubenböck, H. Rapid domain adaptation for disaster impact assessment: Remote sensing of building damage after the 2021 Germany floods. *Sci. Remote. Sens.* **2025**, *12*, 100287. [CrossRef]

5. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25, pp. 1097–1105.
7. 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. [[CrossRef](#)]
8. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241. [[CrossRef](#)]
9. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. . [[CrossRef](#)]
10. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851. [[CrossRef](#)]
11. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890. [[CrossRef](#)]
12. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856. [[CrossRef](#)]
13. 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. [[CrossRef](#)]
14. Wang, J.; Chen, T.; Zheng, L.; Tie, J.; Zhang, Y.; Chen, P.; Luo, Z.; Song, Q. A multi-scale remote sensing semantic segmentation model with boundary enhancement based on UNetFormer. *Sci. Rep.* **2025**, *15*, 14737. . [[CrossRef](#)] [[PubMed](#)]
15. Ma, Q.; Liu, H.; Jin, Y.; Liu, X. Multi-scale context enhancement network with local–global synergy modeling strategy for semantic segmentation on remote sensing images. *Electronics* **2025**, *14*, 2526. [[CrossRef](#)]
16. Cao, Y.; Liu, C.; Wu, Z.; Zhang, L.; Yang, L. Remote sensing image segmentation using Vision Mamba and multi-scale multi-frequency feature fusion. *Remote. Sens.* **2025**, *17*, 1390. [[CrossRef](#)]
17. 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](#)]
18. Zhu, Q.; Li, H.; He, L.; Fan, L. SwinMamba: A hybrid local–global mamba framework for enhancing semantic segmentation of remotely sensed images. *Digit. Signal Process.* **2026**, *175*, 106029. [[CrossRef](#)]
19. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote. Sens.* **2021**, *181*, 84–98. [[CrossRef](#)]
20. 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](#)]
21. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7262–7272. [[CrossRef](#)]
22. 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](#)]
23. Wu, H.; Huang, P.; Zhang, M.; Tang, W.; Yu, X. CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.* **2023**, *61*, 2004612. [[CrossRef](#)]
24. Ma, X.; Zhang, X.; Pun, M.O. RS3Mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote. Sens. Lett.* **2024**, *21*, 6011405. [[CrossRef](#)]

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