



CW-SwinUNet: a novel semantic segmentation approach for very-high-resolution remote sensing imagery

Zhanyuan Chang, Mingyu Xu, Yuwen Wei & Jie Lian

To cite this article: Zhanyuan Chang, Mingyu Xu, Yuwen Wei & Jie Lian (2025) CW-SwinUNet: a novel semantic segmentation approach for very-high-resolution remote sensing imagery, International Journal of Remote Sensing, 46:22, 8614-8639, DOI: [10.1080/01431161.2025.2571233](https://doi.org/10.1080/01431161.2025.2571233)

To link to this article: <https://doi.org/10.1080/01431161.2025.2571233>



Published online: 21 Oct 2025.



Submit your article to this journal [↗](#)



Article views: 61



View related articles [↗](#)



View Crossmark data [↗](#)



CW-SwinUNet: a novel semantic segmentation approach for very-high-resolution remote sensing imagery

Zhanyuan Chang, Mingyu Xu, Yuwen Wei and Jie Lian

College of Information, Mechanical and Electrical Engineering, Shanghai Normal University Shanghai, China

ABSTRACT

The use of deep neural networks for semantic segmentation of RS (remote sensing) images is an important research topic in the field of remote sensing intelligent interpretation, which has significant application value in urban planning, disaster assessment, carbon estimation, and other fields. However, high-resolution RS images bring problems of large data scale, high computational complexity, diverse scales, and irregular shapes. Moreover, there may be distant spatial relationships between target objects, requiring the utilization of long-range contextual information for accurate semantic segmentation. Inspired by the powerful global modelling capability of the Swin Transformer, we propose a new RS image semantic segmentation network. Swin Transformer is used as an encoder to extract global features. To prevent the loss of local details in the encoding stage, one LKA-PM (LKA-Patch Merging) module is designed to extract local detail information by decomposing large kernel convolutions, thereby enhancing the feature representation ability of small objects. Furthermore, to decode global and local features and obtain multi-scale information, one WORC (Weighted Operation Re-parameterization Convolution) module is designed to improve the adaptability and accuracy of the model to different RS images. Finally, to preserve spatial details during decoding and reduce the problem of boundary blurring caused by occlusion in RS imagery, the Coordswin module is designed to provide coordinate information and improve the discrimination ability of similar objects. Experimental results demonstrate that our CW-SwinUNet achieves excellent segmentation performance (91.8 % F1 and 85.2 % MIoU) on the Vaihingen dataset and (93.8 % F1 and 87.9 % MIoU) on Potsdam datasets, surpassing other state-of-the-art methods. The source code of the proposed CW-SwinUNet is available at <https://github.com/xmy1135/cswin/tree/main>.

ARTICLE HISTORY

Received 17 June 2025

Accepted 1 October 2025

KEYWORDS

High-resolution RS imagery; Swin Transformer; reparameterization module; Global Attention Network Model; semantic segmentation

1. Introduction

With the continuous advancement of remote sensing imaging technology, the resolution of acquired images has steadily increased, providing abundant geospatial information. Efficient extraction of key ground object information from these images has become a central task in current earth observation missions (T. Liu et al. 2024). High-resolution

remote sensing images are widely used for semantic segmentation, classification, and pixel-level analysis (K. Liu et al. 2025), supporting numerous applications such as ground object change detection (Miao et al. 2024), land cover classification (Y. Li et al. 2022), and urban planning and monitoring (Z. Chen et al. 2022). In recent years, deep learning techniques have become mainstream for extracting relevant information from remote sensing images and performing semantic segmentation, especially convolutional neural networks (CNNs). Semantic segmentation of remote sensing imagery is of great research value and has broad applications (Hu et al. 2024). However, remote sensing images contain diverse ground objects, which vary significantly in size, shape, and spectral characteristics. For example, buildings, grasslands, and shrubs exhibit multi-scale features, and their spatial arrangements are often irregular, resulting in complex spatial relationships (Jing et al. 2025). Optimizing semantic segmentation algorithms to accurately capture these complex spatial patterns remains a significant challenge. To address this issue, CW-SwinUnet (Hu et al. 2024), a novel method based on the Swin Transformer, was proposed. This method improves segmentation performance across different object scales and enhances the discrimination of small targets, thereby effectively increasing overall segmentation accuracy.

In remote sensing earth observation tasks, CNNs have driven the development of semantic segmentation; however, their limited receptive fields only allow them to capture local information. Although efforts have been made to obtain global semantic information by stacking multiple convolutional layers and performing downsampling operations, this often results in the loss of fine-grained features (X. Wang et al. 2024). To mitigate the loss of global contextual features during feature extraction, extensive studies have focused on the design and optimization of neural network architectures. For instance, Li et al. (2022) proposed a Collaborative Boosting Framework (CBF), which iteratively combines deep learning modules with knowledge-guided ontological reasoning modules to enhance the model's discriminative capability for spatial information in remote sensing images, addressing issues of coarse segmentation and the lack of prior knowledge guidance. However, its overly simplified decoder leads to coarse-resolution segmentation, limiting the accuracy of semantic segmentation. To overcome this limitation, Zhu et al. (2025) proposed the UNetMamba model based on Mamba, in which the MSD module efficiently decodes semantic information from multi-scale feature maps, and a Local Supervision Module (LSM) is introduced to enhance the perception of local semantic information. Sha et al. (2025) proposed a multi-optimization strategy for reconstructing 3D models of transparent objects under limited constraints, extracting and modelling features through normal vectors, followed by weight updates to reduce convergence issues. Based on the encoder-decoder architecture, various techniques have been introduced to enhance semantic segmentation in remote sensing. For example, Zhang et al. (2023) proposed an improved U-Net model based on transfer learning, incorporating a transfer learning mechanism in the encoder and performing multi-scale fusion by bilinear upsampling followed by sequential integration with the corresponding encoder feature maps in the decoder. Zhou et al. (Jin et al. 2022) proposed DenseUNet, which introduces dense connections at each decoder level to facilitate feature propagation and reuse within the network. Xu et al. proposed a novel three-branch network, PIDNet (Xu, Xiong, and Bhattacharyya 2023), in which each branch is responsible for parsing detail information, contextual information, and boundary information, respectively. Ding et al.

(2024) proposed UniRepLKNet, which employs parallel large convolutional kernels and dilated convolutions to maintain the receptive field while transforming low-level features into high-level abstract features. Nevertheless, due to the presence of small targets, high similarity among objects, complex target distributions, and occlusions between targets, semantic segmentation of high-resolution remote sensing images still faces new challenges. Cai et al. (2024) proposed PKINet, which extracts dense texture features under different receptive fields through multi-scale parallel depthwise separable convolutions, thereby aggregating local contextual information. Additionally, it introduces a Contextual Anchor Attention mechanism to capture long-range contextual information. During feature extraction, CNN models typically perform feature downsampling to reduce computational complexity, which may lead to the loss of local details and the neglect of small target features (Guo et al. 2023). Different land cover categories may share similar feature information, making it difficult for CNNs to distinguish them (H. Feng et al. 2025). Furthermore, CNNs inherently lack the ability to model global contextual information or long-term dependencies, which may result in segmentation errors caused by occlusions among targets. Therefore, in semantic segmentation tasks, it is necessary to incorporate both global and local contextual information, as well as additional local fine-grained features as cues for accurate segmentation.

With the advancement of technology, attention modules are increasingly employed to enhance channel interactions and network representation. For example, Li et al. (2022) proposed a Multi-Attention Network (MANet), which extracts contextual dependencies through multiple efficient attention modules and integrates the extracted local feature maps with their corresponding global dependencies. Meng et al. (2025) proposed a Dual-Level Network (DLNet), which captures long-range dependencies via self-attention to enhance contextual understanding and then leverages cross-attention to facilitate interactions between local and global features, enabling multi-scale feature extraction and fusion. Wang et al. (2024) proposed an improved model based on TransDeepLab, which introduces a GAM attention mechanism in the encoding stage and employs a multi-level linear upsampling strategy in the decoding stage, thereby enhancing the capture of multi-level semantic information and small target details. Feng et al. (2025) proposed a hyper-rectangle embedding method based on 3D scene prediction, employing progressive formation and attention-based information diffusion models to guide the network in generating more balanced information for fine-grained and intuitive entity modelling. In addition, multi-scale feature fusion strategies have been widely adopted. For instance, Zhu et al. (2025) proposed a Global-Local Feature Fusion Network (GLFFNet), which uses a Residual Network as the main branch to extract local features and introduces VMamba as an auxiliary branch encoder to provide global information to the main branch. Chen et al. (2025) proposed an enhanced AE-UNet++ network, in which the Context Feature Fusion (CFF) module and the Spatial Awareness Fusion (SAF) module effectively work together to improve the model's ability to distinguish buildings from complex backgrounds. Wang et al. (2025) proposed a Lightweight Multimodal Fusion Network (LMFNet), where the multimodal feature fusion reconstruction layer and the multimodal feature self-attention fusion layer are effectively combined to enable the reconstruction and fusion of multimodal features. Sha et al. (2025) proposed a multi-task joint learning network (SSC-NET), which extracts features through an ROI feature module and subsequently performs feature fusion. Ma et al. (2025) proposed a Multi-Scale Spatio-Temporal

Interaction Fusion Network (MSIFN), which highlights key features through a channel attention mechanism and spatial convolution while performing spatial modelling, and finally fuses spatial and temporal features via a gated recurrent unit-temporal convolution network. Xiang et al. (2025) proposed an unsupervised underwater image restoration network based on homogeneous constraints for multi-parameter estimation, employing a residual network with contextual attention to estimate the scene while integrating local and global features. Although the aforementioned methods effectively capture contextual information, they still exhibit omissions in small objects and boundary details.

In recent years, with Dosovitskiy introducing the Transformer to image classification tasks, the Vision Transformer (ViT) model was proposed. This model is entirely built upon the self-attention mechanism and benefits from a global receptive field. Unlike traditional CNN architectures, the success of Transformer enables the extraction of global information. By transforming tasks based on 2D images into 1D sequence-based tasks, transformer fully exploits its powerful sequence modelling capability, demonstrating excellent global feature extraction ability. Inspired by Transformer, Chen et al. developed the Swin Transformer, which has shown great potential in various dense prediction tasks. Subsequently, Wang et al. proposed the Unetformer (L. Wang, Li, Zhang, et al. 2022) network for real-time urban scene segmentation in remote sensing images. They introduced an effective global-local attention mechanism (GLTB) and developed a lightweight Transformer-based decoder, which significantly enhanced the network's ability to extract multi-scale contextual features, thereby improving remote sensing semantic segmentation performance. In the same year, they also proposed ST-Unet (He et al. 2022), a novel dual-encoder architecture consisting of Swin Transformer and CNN in parallel. By encoding spatial information through a spatial interaction module (SIM), pixel-level correlations are established within the Swin Transformer blocks, enhancing the feature representation of occluded objects. Next, Tian et al. proposed GLFFNet (Tian et al. 2023), a network based on Transformer, convolution, and their variants, which designs two high-performance encoders to extract global high-order interaction features and low-order local features, thereby improving the recognition of small targets. In the same year, Chen et al. proposed a new convolution-transformer hybrid semantic segmentation model, CTFuse (X. Zhu et al. 2022), which incorporates spatial and channel attention modules into the Transformer to enhance both global and local feature representations. Xia et al. subsequently proposed ViT-CoMer (Xia et al. 2024), which combines Vision Transformer (ViT) with multi-scale convolutions to strengthen feature interaction and diversity. By employing a CNN-Transformer bidirectional interaction module, this approach enriches features from both architectures and facilitates cross-layer multi-scale feature fusion. Wang et al. proposed MCAT-UNet (T. Wang et al. 2024), which integrates multi-scale convolutional attention with a cross-shaped window self-attention mechanism, effectively modelling long-range spatial dependencies while extracting local feature representations, enabling more precise multi-scale object detection. Zhang et al. proposed ConvLSR-Net (R. Zhang, Zhang, and Zhang 2024), which appends Transformer modules after each CNN stage in a pure convolutional network to extract multi-scale information, allowing the model to learn both local and global representations at each stage. Zhou et al. proposed MSGCNet (Zeng et al. 2024), which applies efficient multi-head self-attention (MSA) and channel attention within local windows of the Swin Transformer to reduce computational complexity while incorporating spatial attention and global context interaction in the

decoder, thereby enhancing the model's capacity to capture global information. Overall, Transformer-based methods effectively capture global contextual information but incur high computational complexity. Particularly when processing ultra-high-resolution remote sensing images, these methods face challenges in balancing accuracy and computational efficiency.

To address the aforementioned issues, we propose a novel network named CW-SwinUnet, which employs a Swin Transformer as the encoder and a dual-branch decoder based on Transformer and CNN. Specifically, the Swin Transformer encoder extracts global feature information, while the LKA-PM module is designed to preserve local details during encoding. The dual-branch decoder extracts global contextual information and fine-grained details at different scales, enabling comprehensive decoding of both global and local features. Additionally, we design the WORC module to improve the model's adaptability and accuracy across diverse remote sensing images without increasing computational cost. Finally, the Coordswin module preserves spatial information and reduces edge blurring caused by occlusion in remote sensing images. The main contributions of this work are summarized as follows:

- (1) We construct the LKA-PM module, which focuses on capturing local details during encoding, collecting features of small objects, and avoiding loss of local information, thereby improving segmentation of small targets.
- (2) To enhance adaptability and accuracy across various remote sensing images without extra computational cost, we design the WORC module. This module improves segmentation accuracy and detail preservation, effectively distinguishing highly similar ground objects.
- (3) We propose the Coordswin module, which emphasizes pixel-level spatial feature correlations, preserves more spatial information, and reduces edge blurring caused by occlusion in remote sensing images.

2 Materials and methods

2.1. Materials

To evaluate the proposed method, this paper performs a series of experiments on the ISPRS Potsdam and Vaihingen datasets, comparing its performance with that of other methods.

2.1.1. Vaihingen datasets

As shown in Figure 1, the Vaihingen dataset is a benchmark dataset for semantic segmentation of remote sensing (RS) imagery, containing 33 true orthophotos (TOP) collected from advanced airborne sensors, covering a total area of 1.38 km² in the Vaihingen region. The images in the dataset have a very high spatial resolution, with an average size of 2494 × 2064 pixels and a ground sampling distance (GSD) of approximately 9 cm. Each TOP image block contains three multispectral bands (near-infrared, red and green) as well as a digital surface model (DSM) and a normalized digital surface model (NDSM). The dataset involves five foreground classes, including impervious surfaces, buildings, low vegetation, trees, and cars, as well as a background class (clutter). In the

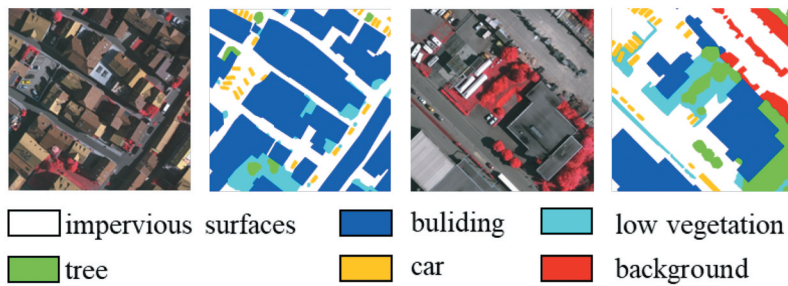


Figure 1. Partial example plot of the Vaihingen dataset.

experiments, we followed (L. Wang, Li, Zhang, et al. 2022; T. Wang et al. 2024) to select 11 images for training (image IDs: 1, 3, 5, 7, 13, 17, 21, 23, 26, 32 and 37), and the remaining images for testing. In the experiments, we only used the red, green, and blue channels and the true labels with eroded boundaries. The image blocks were cropped into smaller patches of 1024×1024 pixels.

2.1.2. Potsdam datasets

The Potsdam dataset is a benchmark dataset for semantic segmentation of remote sensing (RS) imagery, an example is shown in Figure 2. It consists of 38 image patches of the same size (6000×6000 pixels) (Rottensteiner et al. 2014). These patches are extracted from high-resolution true orthophoto (TOP) mosaics with a ground sampling distance (GSD) of 5 cm. The dataset covers a 3.42 square kilometre area in the city of Potsdam, which is characterized by complex buildings and dense urban structures. The dataset is labelled with six categories for semantic segmentation studies. Each image provides three channel combinations: infrared-red-green (IR-R-G), red-green-blue (R-G-B), and red-green-blue-infrared (R-G-B-IR). Each image patch has a size of 6000×6000 pixels, and the class information in the dataset is the same as the Vaihingen dataset. Each image patch provides four multispectral bands (red, green, blue, and near-infrared) as well as a digital surface model (DSM) and a normalized digital surface model (NDSM). To ensure fair comparative experimental conditions, we followed previous works (L. Wang, Li, Zhang, et al. 2022; T. Wang et al. 2024) The image patches with 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15 and 7_13 were used for testing, while

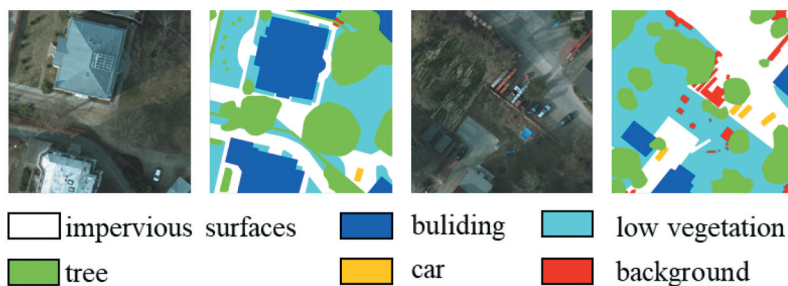


Figure 2. Partial example plot of the Potsdam dataset.

the remaining 23 patches were used for training. Similarly, only the red, green, and blue channels were used in the experiments, and the original image patches were cropped into blocks of 1024×1024 pixels.

1.2. Methods

In this section, we first introduce the overall structure of CW-SwinUnet and describe the standard Swin Transformer module. After that, we discuss three important modules in CW-SwinUnet, namely LKA-PM, WORC (Weighted Online Re-parameterization Convolution), and CoordSwin.

The overall architecture of our CW-SwinUnet is illustrated in Figure 3. As a hybrid of the SwinUnet decoder module and CNN, CW-SwinUnet adopts the effective architecture of SwinUnet. In this structure, we use skip connections to connect the encoder and decoder. Specifically, CW-SwinUnet incorporates a CNN-based local feature extraction module and a Swin Transformer-based global feature extraction module. By utilizing the LKA-PM module, feature information is efficiently propagated, enabling comprehensive extraction of global features in RS imagery. Additionally, the performance of the Swin Transformer decoder module is enhanced through the designed WORC and CoordSwin modules. For

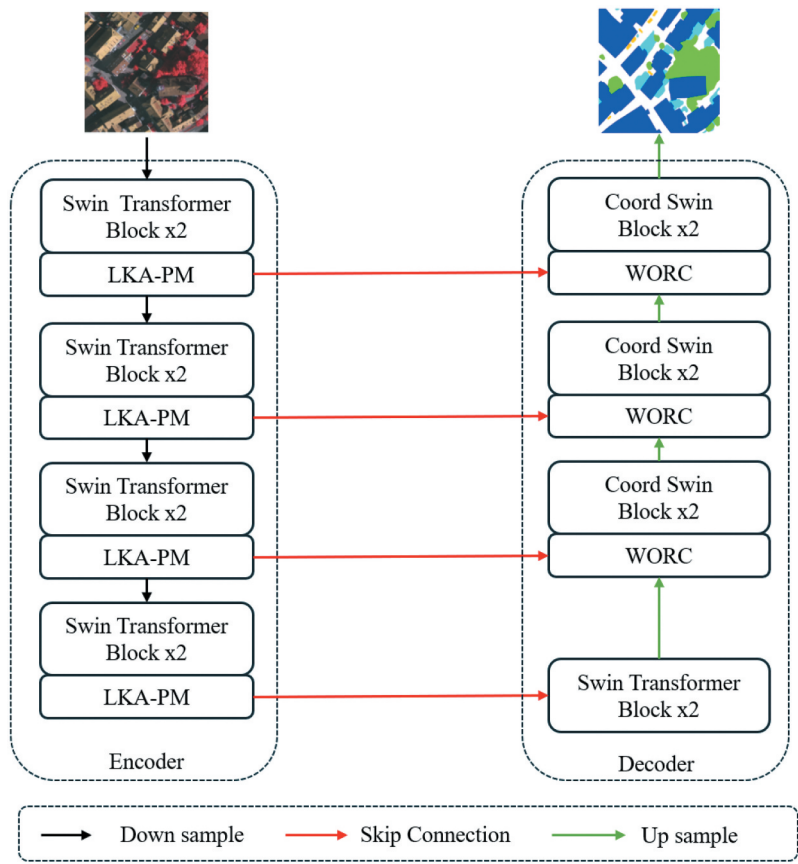


Figure 3. The overall architecture of CW-SwinUnet.

an input remote sensing image $X \in \mathbb{R}^{H \times W \times 3}$, the Swin Transformer encoder divides the image data into non-overlapping small patches using a sliding window approach. Within each window, self-attention computations are performed, capturing long-range dependencies across the global scope through hierarchical self-attention calculations. However, due to the window-based self-attention mechanism employed by Swin Transformer, the window size determines the receptive field for each position. Smaller window sizes may result in the loss of local details, especially for tasks involving fine-grained structures. Additionally, compared to traditional convolutional neural networks, Swin Transformer typically has a larger number of parameters, increasing the training and inference costs of the model. To avoid losing local details during the encoding stage while reducing the training cost of the model, we propose the LKA-PM module. The LKA-PM module performs downsampling before each stage to reduce the computational complexity while preserving local feature information. In the encoder, the 3D tokenized input with a resolution of $H \times W$ is fed into two consecutive Swin Transformer blocks for representation learning. In our experiments, the window size and the number of attention heads were set to 8, while the feature dimension and resolution remained unchanged. Simultaneously, the LKA-PM module (Large Kernel Convolutional Patch Merging layer) downsamples the feature resolution by a factor of 2 (2x downsampling) and increases the feature dimension to twice the original dimension. This process is repeated three times in the encoder. After the encoding stage described above, we obtain a set of feature representations: $F = R(H/2) \times (W/2) \times 64$. The features obtained from the encoding stage are input into the Residual Adaptive CoordConv (RAC) decoder. The decoder consists of three CoordSwin modules and a WORC module, as shown in the diagram. Unlike the standard Swin Transformer modules used in the encoder, the CoordSwin modules in the decoder improve upon the standard Swin Transformer by incorporating the CoordConv module and fusing it with CNN feature extraction modules. Similarly, instead of using patch merging layers as in the encoder, we use WORC layers in the decoder to upsample the extracted deep features. The WORC layer reshapes adjacent-dimensional feature maps into higher-resolution feature maps (2x upsampling) while reducing the feature dimension to half of the original dimension. Similar to U-Net, skip connection layers are used to connect the encoder and decoder, and the channel numbers passing through the skip connection layers are reduced using 3×3 convolutional layers. Additionally, each WORC module is accompanied by an interpolation module and a ReLU layer. The above process is repeated four times, gradually expanding the features F . Finally, the features are subjected to post-processing using 1×1 convolutional layers and upsampling operations to generate the final segmentation map.

1.2.1. Swin transformer block

Consistent with what was mentioned earlier, the Swin Transformer block is built based on the shifted windows, which differs from the traditional multi-head self-attention (MSA) module. Each Swin Transformer block consists of a LayerNorm (LN) layer, a multi-head self-attention module, residual connections, and a 2-layer MLP with GELU nonlinearity. Specifically, the Swin Transformer module adopts two different multi-head self-attention algorithms: window-based multi-head self-attention (W-MSA) and shifted window-based multi-head self-attention (SW-MSA) to replace MSA. As illustrated in Figure 4, W-MSA and

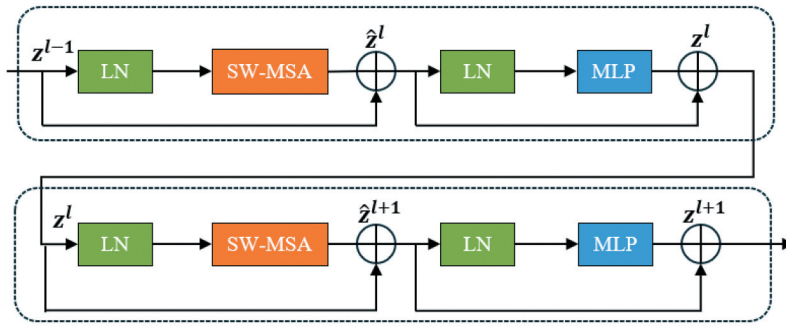


Figure 4. Swin Transformer module structure.

SW-MSA alternate between two consecutive Swin Transformer blocks. The output of the l -th block z^l and \hat{z}^l can be expressed as:

$$\hat{z}^l = W - \text{MSA}(\text{LN}(z^{l-1})) + z^{l-1} \quad (1)$$

$$z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \quad (2)$$

$$\hat{z}^{l+1} = \text{SW} - \text{MSA}(\text{LN}(z^l)) + z^l \quad (3)$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \quad (4)$$

The formula for calculating self-attention is as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (5)$$

Where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ represent the query, key, and value matrices. M^2 and d , respectively, denote the number of patches in the window and the dimension of the query or key. Moreover, the values of the matrix B are taken from the bias matrix $B = \mathbb{R}^{(2M-1) \times (2M+1)}$.

1.2.2. LKA-PM module

Details and edge information in RS imagery are crucial for semantic segmentation. Due to the windowed self-attention mechanism employed by Swin Transformer, the window size determines the perception range of each position. Larger window sizes may lead to the loss of local details, especially for image tasks with fine-grained structures, where local details may not be fully captured. Additionally, RS imagery typically has large spatial coverage and high resolution, posing challenges in terms of computational complexity and memory consumption. Compared to traditional convolutional neural networks, Swin Transformer usually has a larger number of parameters, increasing the training and inference costs of the model. To avoid losing local details in the encoding stage while reducing the training cost of the model, we propose the Patch Merging module by incorporating LKA attention (Lau, Po, and Rehman 2023), where the components of LKA-PM are illustrated in Figure 5.

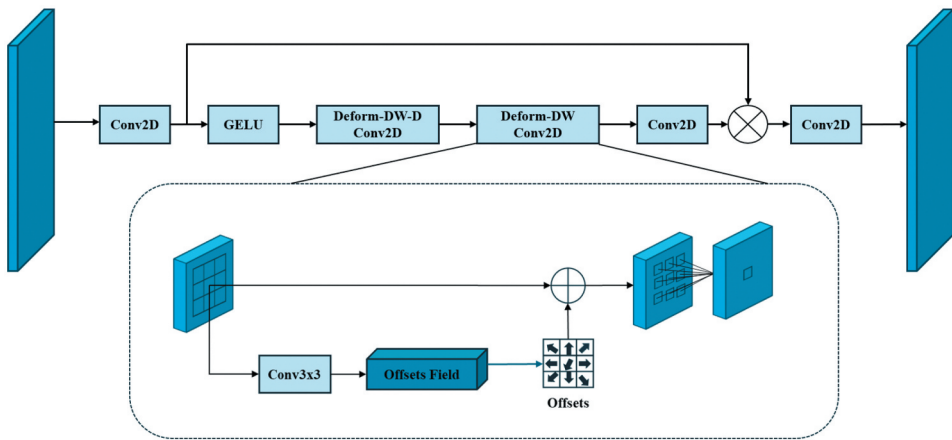


Figure 5. LKA-PM module structure.

Due to the decomposition of large kernel convolutions, LKA-PM can capture long-range relationships and local details simultaneously, enhancing small target object segmentation capability without increasing the number of parameters. Specifically, we can decompose a large kernel convolution using three different types of convolutions: an expandable spatial local convolution that focuses on nearby pixel interactions and offers detailed local context, a depth spatial convolution that processes interactions over longer distances within the feature map and includes broader contextual information and a channel convolution (1×1 convolution) that reduces dimensionality and efficiently blends information across various feature channels. The LKA module formula is as follows:

$$\text{Attention} = \text{Conv}_{1 \times 1}(\text{DW} - \text{D} - \text{Conv}(\text{DW} - \text{Conv}(F))) \quad (6)$$

$$\text{Output} = \text{Attention} \otimes F. \quad (7)$$

Here, $F \in \mathbb{R}^{C \times H \times W}$ denotes the input features. $\text{Attention} \in \mathbb{R}^{C \times H \times W}$ represents the attention map. The values in the attention map indicate the importance of each feature. \otimes represents an element-wise product.

Figure 6 shows the decomposition of large kernel convolutions into depth convolution, depth dilation convolution, and pointwise convolution, emphasizing LKA's capacity to integrate convolution and self-attention benefits. We apply a 5×5 depth convolution to the input x , yielding an initial attention map that is subsequently processed with a 7×7

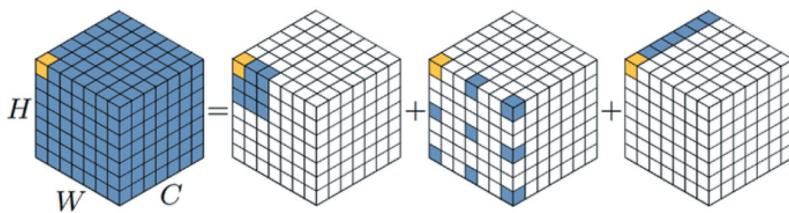


Figure 6. Schematic diagram of LKA-PM.

spatial convolution to create a higher-level attention matrix. The attention map is adjusted using a 1×1 convolution to produce the final attention weights. Finally, the input x is element-wise multiplied by these attention weights, resulting in a weighted feature map. In the figure, the centre point is shown to be a yellow grid, and the convolution positions are marked by blue grids.

1.2.3. Coordswin module

RS imagery often has large scales and complex spatial structures. Coordinate information enhances the model’s ability to detect variations in features across different locations. In semantic segmentation tasks, certain classes may be more common in specific geographic locations. Additionally, Swin Transformer blocks establish relationships between patch tokens within a limited window, effectively reducing memory overhead. However, this approach somewhat weakens the global modelling capability of the Transformer even with the alternating execution strategy of regular and shifted windows. Moreover, occlusions of objects in RS imagery can cause blurred boundaries, requiring spatial information for elimination. Therefore, as depicted in Figure 7, we propose applying CoordConv to the MLP layer of Swin Transformer. CoordConv helps the model capture spatial correlations, enhancing the model’s spatial position perception ability in semantic segmentation tasks for RS imagery. This improves the model’s handling of features at different scales and resolutions while alleviating the limitations of translation invariance. These advantages contribute to improving the accuracy and robustness of semantic segmentation in RS imagery, enabling the model to better understand and interpret spatial information in RS imagery.

The CoordConv layer enhances standard 2D convolution by adding channels with hard-coded i and j coordinates to the input representation. It introduces additional channels to the input data to enrich the spatial representation.

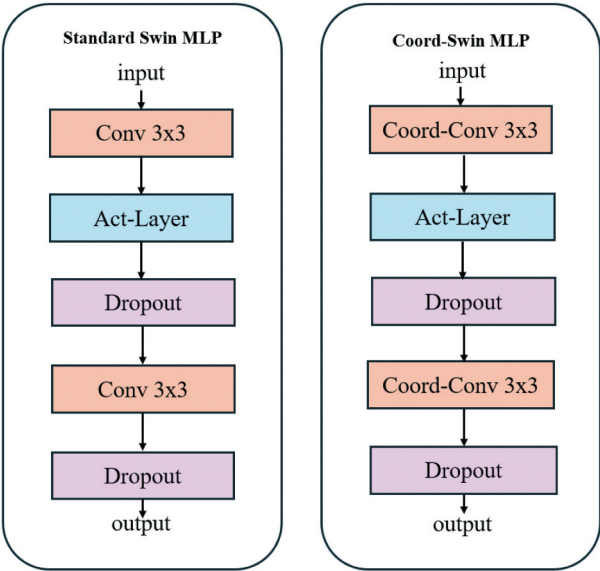


Figure 7. Coordswin module structure.

Specifically, as illustrated in Figure 8, CoordConv uses predefined i and j coordinates to provide explicit spatial information. The i -coordinate is a $h \times w$ rank-1 matrix representing the row indices of the input matrix. The j -coordinate represents the column indices of the input matrix, with rows and columns in the matrix filled with 0s and 1s, respectively. This enhances the performance and capability of standard convolutional layers by incorporating detailed positional data into the input.

1.2.4. WORC module

RS imagery often contains multi-scale geospatial information, such as small-scale buildings and large-scale land cover. In previous feature extraction modules, introducing additional feature extractors can capture richer scale information and fuse them in the decoder. The reparameterization module can adaptively learn the weights for feature fusion to combine multi-scale features optimally, thereby improving the accuracy and preservation of fine details in RS imagery semantic segmentation. Additionally, due to the complex distribution of objects and diverse types of land cover in RS imagery, introducing additional feature extractors allows the model to better adapt to different types of geospatial features. The reparameterization module adaptively adjusts the fusion weights according to the input data characteristics, thereby improving the adaptability and accuracy of the model for different RS imagery. To address this, we propose a WORC module, which processes the fused features using a reparameterization module. The WORC module is illustrated in Figure 9.

The intermediate normalization layers are crucial components of the multi-layer and multi-branch structures in reparameterization. Studies have shown that removing these layers leads to significant performance degradation, highlighting their importance for the performance of the reparameterized model. The use of intermediate normalization layers unexpectedly results in a higher training cost. However, the reparameterization module can adaptively learn the weights for feature fusion optimally by combining multi-scale features, improving the accuracy and detail preservation capabilities of semantic segmentation in RS imagery. Specifically, as depicted in Figure 10, we complete re-parameterization in three steps. First, we remove all batch normalization layers. Second, scaling layers are

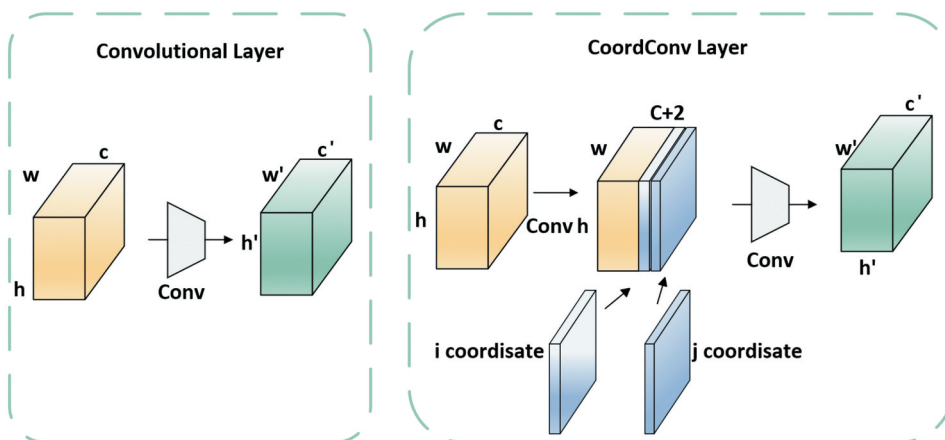


Figure 8. Coord-conv structure.

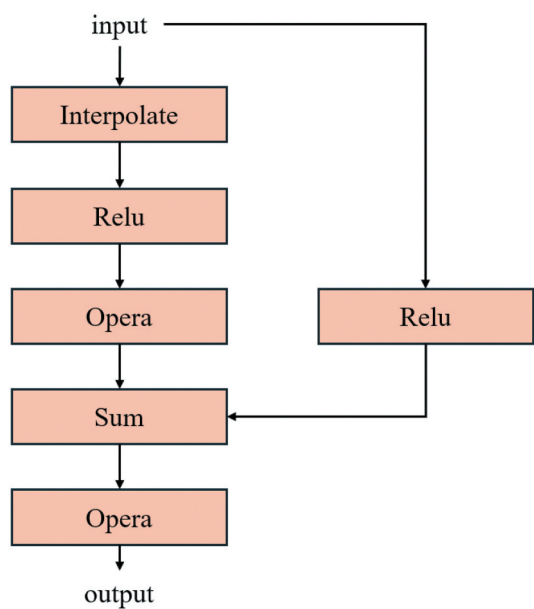


Figure 9. WORC module structure.

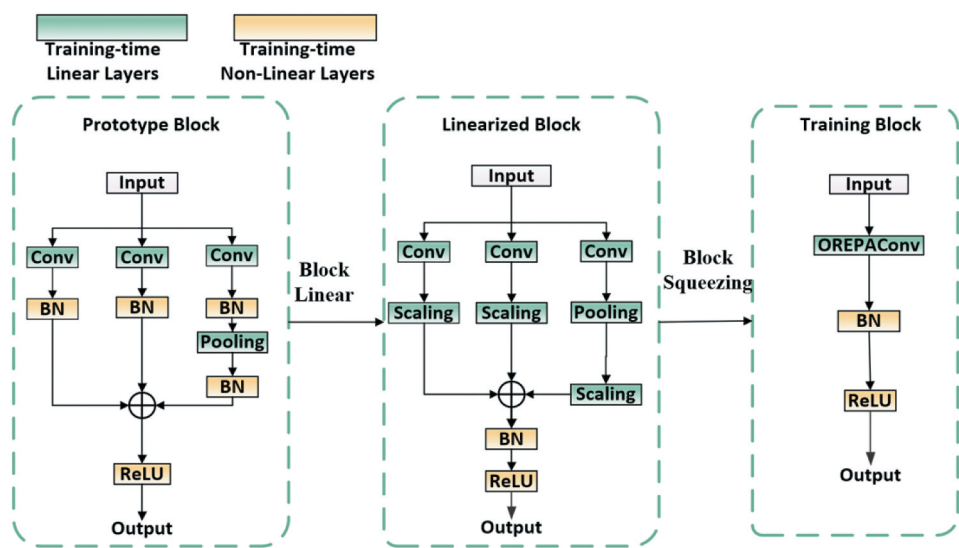


Figure 10. Opera module structure.

added to maintain multiple optimization paths to enhance model robustness. Finally, post-addition normalization is performed to ensure training stability, simplifying the re-parameterization block to only linear layers during training, ensuring that only linear components remain after the linearization stage and allowing for the merging of all linear components during training to enhance efficiency.

2. Experiments and results

This section validates our proposed method through comparative experiments and ablation studies, describing evaluation metrics, network parameters, and experimental setup. We evaluated the relative strengths and weaknesses of our method compared to other SOTA (state-of-the-art) methods through comparative experiments and conducted ablation studies to assess the contributions of different components and configurations in the network.

2.1. Implementation details

2.1.1. Training settings

During training, we employed random vertical flipping, random horizontal flipping, and random Gaussian erasing as data augmentation strategies. For testing, we utilized multi-scale and random flipping data augmentation strategies. Our network was implemented using the PyTorch framework. To ensure rapid convergence, we utilized the Adam optimizer for training all models. The initial learning rate was set to $6e-4$, adjusted using a cosine annealing strategy. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU (24 GB), with a batch size of 8 and a maximum of 100 epochs.

2.1.2. Loss function

The class proportions in the Vaihingen and Potsdam datasets are imbalanced, causing the model to focus more on classes with larger proportions while ignoring those with smaller proportions. To address this issue, we employed a combination of Dice loss (L_{Dice}) and cross-entropy loss (L_{CE}) as the supervised loss functions. The joint loss (L) is represented as follows:

$$L = L_{\text{CE}} + L_{\text{Dice}} \quad (8)$$

$$L_{\text{CE}} = -\frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K y_k^{(n)} \log_c \hat{y}_k^{(n)} \quad (9)$$

$$L_{\text{Dice}} = 1 - \frac{2}{N} \sum_{n=1}^N \sum_{k=1}^K \frac{\hat{y}_k^{(n)} y_k^{(n)}}{\hat{y}_k^{(n)} + y_k^{(n)}} \quad (10)$$

2.1.3. Evaluation index

We use two commonly used semantic segmentation evaluation metrics to assess the performance of the model, including Intersection over Union (IoU) and average F1 (Ave. F1). Generally, there are four relationships between the predicted results and the true classes: (1) TP (True Positive): positive samples that are correctly predicted; (2) FP (False Positive): negative samples that are incorrectly predicted as positive; (3) TN (True Negative): negative samples that are correctly predicted; (4) FN (False Negative): positive samples that are incorrectly predicted as negative. The Intersection over Union (IoU) for each category is defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{FP} + \text{FN} + \text{TP}} \quad (11)$$

The F1 score represents the harmonic mean of Precision and Recall, and is defined as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

Here, precision denotes the proportion of true positive samples among all samples predicted as positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

Recall denotes the proportion of true positive samples correctly predicted among all actual positive samples:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

In addition, MIoU represents the mean Intersection over Union across all classes, while MF1 represents the mean F1 score across all classes.

2.2. Comparative experiments

To evaluate the segmentation performance of the model proposed in this paper, we compared it with other commonly used remote sensing image semantic segmentation models, including MANet (G. Zhu and Kim 2021), SwiftNet (H. Wang et al. 2021), SwinUperNet (Z. Wang et al. 2023), DeepLabV3+ (L.-C. Chen et al. 2018), FTUNetFormer (Tian et al. 2023), DC-Swin (L. Wang, Li, Duan, et al. 2022), MCAT-UNet (T. Wang et al. 2024), DEDNet (Z. Wang et al. 2024). To ensure fairness, all methods were tested using the same code, and the results are highlighted in bold in the table. These tables provide clear evidence that the experimental results of CW-SwinUNet are superior to classical semantic segmentation networks, such as DeepLabV3+. Nonetheless, the segmentation performance of SwinUnet decreases when facing highly similar patches and small target patches even with Swin Transformer-based encoders and decoders. Therefore, we enhance SwinUNet by integrating the LKA-PM module, WORC reparameterization module, and CoordSwin module. This approach effectively improves the segmentation performance for objects of different scales, elongated objects, and object boundaries.

Experimental results on the Potsdam dataset are shown in Table 1. CW-SwinUnet surpasses recent Transformer-based networks like UNetFormer and CNN-based MANet. CW-SwinUnet demonstrates strong performance across all categories, achieving an average F1 score of 93.8 % and an MIoU score of 87.9 %. Due to the different data types, the segmentation accuracy on the Potsdam dataset is generally higher than that on the Vaihingen dataset. It is worth noting that the proposed CW-SwinUnet still outperforms other CNN and Transformer-based networks. Among them, the hybrid CNN and Transformer network, FTUNetFormer, surpasses other CNN-based models, indicating that the combination of Transformer's global modeling ability and CNN's local feature extraction ability helps semantic segmentation models obtain more accurate and powerful segmentation results. Figure 11 displays

Table 1. Segmentation accuracy of different methods on the Potsdam dataset.

Model	Backbone	F1(%)					Evaluation index	
		Surface	Building	Low	Tree	Car	MF1	MIoU
MANet	ResNet50	93.4	97.0	88.3	89.4	96.5	92.9	87.0
SwiftNet	ResNet50	91.8	95.9	85.7	86.8	94.5	91.0	83.8
DeepLabV3+	ResNet50	92.1	95.3	85.6	86.5	94.8	90.9	84.2
SwinUperNet	Swin-Small	93.2	96.4	87.6	88.6	95.4	92.2	85.8
FTUNetFormer	Swin-Base	93.9	97.2	88.8	88.8	96.6	93.3	87.5
DC-Swin	Swin-Small	93.5	96.9	87.9	89.1	95.8	92.7	86.5
MCAT-UNet	MSCAN-S	94.0	96.4	87.1	88.1	95.9	92.9	86.8
DEDNet	/	94.0	95.4	87.9	86.8	95.9	93.0	86.9
CW-SwinUNet	Swin-Base	94.4	97.4	88.4	89.2	96.8	93.8	87.9

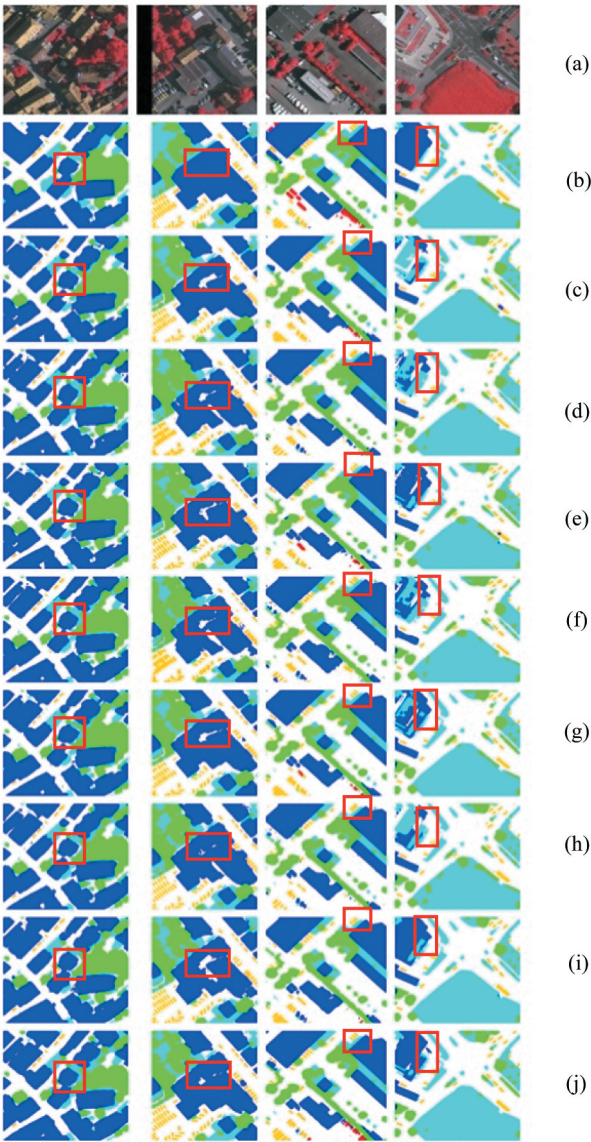


Figure 11. Visual comparison of semantic segmentation for small object features on the Potsdam dataset . (a) Image, (b) Ground truth, (c) DeepLabV3+, (d) SwinUperNet, (e) FTUNetFormer, (f) DC-Swin, (g) MCAT-UNet, (h) DEDNet, (i) MANet, (j) CW-SwinUNet.

Table 2. Segmentation accuracy of different methods on the Vaihingen dataset.

Model	Backbone	F1(%)					Evaluation index	
		Surface	Building	Low	Tree	Car	MF1	MIoU
MANet	ResNet50	93.0	95.5	84.6	90.0	88.9	90.4	82.7
SwiftNet	ResNet50	92.2	94.8	84.1	89.3	81.2	88.3	79.6
DeepLabV3+	ResNet50	91.6	94.1	82.5	88.0	77.7	86.7	77.1
SwinUpNet	Swin-Small	92.8	95.6	85.1	90.6	85.1	89.8	81.8
FTUNetFormer	Swin-Base	93.5	96.0	85.6	90.8	90.4	91.3	84.1
DC-Swin	Swin-Small	93.6	96.2	85.8	90.4	87.6	90.7	83.2
MCA-UNet	MSCAN-S	93.0	95.4	85.3	88.1	88.0	90.5	83.7
DEDNet	/	93.9	94.4	84.9	86.8	87.5	90.8	83.5
CW-SwinUNet	Swin-Base	97.1	96.3	85.3	90.3	90.8	91.8	85.2

the predicted results of the proposed centralized semantic segmentation methods in the table. It can be observed that CW-SwinUnet outperforms Ftunetformer in the segmentation of small target objects, achieving better results. By extracting local details during the encoding process, CW-SwinUnet collects more features of small objects, avoiding the loss of local details and improving the segmentation capability for small targets in images.

Results on the Vaihingen dataset: Table 2 presents the numerical results for each semantic segmentation method. The results indicate that the CW-SwinUnet algorithm outperforms other algorithms in terms of MIoU (85.2 %) and mean F1 (91.8 %), demonstrating clear advantages over both CNN- and Transformer-based networks. The MIoU of CW-SwinUnet is at least 1.1 % higher than other networks. It is worth noting that this method achieves an F1 score of 97.1 % in the surface category, surpassing other networks by more than 3.5 %. Figure 12 shows the predicted results of the proposed centralized semantic segmentation methods. It can be observed that DC-Swin still performs worse than CNN-based MANet, highlighting the importance of spatial coordinate information in remote sensing semantic segmentation. The absence of spatial coordinate information often results in blurred boundaries. Compared to other models, CW-SwinUNet reduces the probability of mis-segmentation by preserving spatial coordinate information, effectively distinguishing highly similar objects, and improving the segmentation capability for small targets in images.

3. Ablation study

Effect of the LKA-PM Module: Based on Table 3, it is evident that the introduction of the LKA-PM module significantly improves the segmentation performance of the baseline Swin-Unet model, particularly for small object segmentation. Specifically, we discuss the segmentation scenarios with and without the LKA-PM module for the usage of the Swin-based encoder. With the LKA-PM module, compared to the baseline, the model achieves a 1.0 % increase in mean F1 and a 0.5 % increase in MIoU for the Potsdam dataset. Particularly in the ‘Car’ category, the LKA-PM module achieves an improvement of 0.8%. The corresponding visual segmentation results are shown in Figure 13.

In the first row of images, it can be observed that the model fails to accurately determine the boundaries of cars and erroneously merges two cars into one (resulting

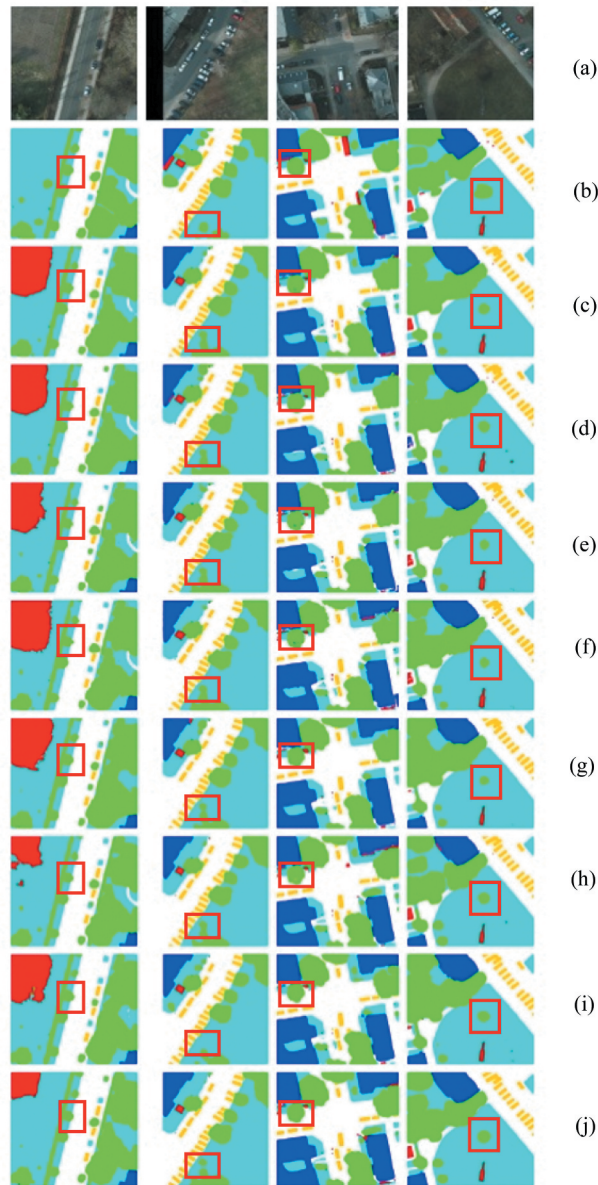


Figure 12. Visual comparison of semantic segmentation for small object features on the Vaihingen dataset. (a) Image, (b) Ground truth, (c) DeepLabV3+, (d) SwinUpperNet, (e) FTUNetFormer, (f) DC-Swin, (g) MCAT-UNet, (h) DEDNet, (i) MANet, (j) CW-SwinUNet.

in boundary blending). However, after incorporating the LKA-PM module, the recognition of car boundaries is enhanced. This provides evidence that decomposing a large kernel convolution operation enables the model to capture more local details, thereby improving the segmentation capability for small objects in images.

Effect of the Coordswin Module: Table 4 demonstrates that considering Coordswin in the CW-SwinUnet framework leads to improvements in segmentation results. Specifically,

Table 3. Ablation experiments of the proposed modules on the Vaihingen dataset.

Model	F1(%)					Evaluation index	
	Surface	Building	Low	Tree	Car	MF1	MIoU
Baseline	92.5	94.9	84.6	90.0	88.9	90.4	82.9
Baseline + LKA-PM	94.1	95.3	85.0	90.1	89.5	91.4	83.4
Baseline + Coordswin	93.1	95.9	84.7	89.5	87.9	90.8	83.9
Baseline + WORC	93.5	95.1	84.9	90.2	89.3	90.7	84.0
Baseline + LKA-PM + WORC	94.5	96.4	85.1	90.5	91.0	91.5	84.9
Baseline + Coordswin + WORC	94.4	96.1	84.7	89.9	89.4	90.9	84.6
Baseline + LKA-PM + Coordswin	96.2	96.0	85.1	89.8	90.3	91.6	85.0
Baseline + LKA-PM+ Coordswin+ WORC	97.1	96.2	85.3	90.3	90.8	91.8	85.2

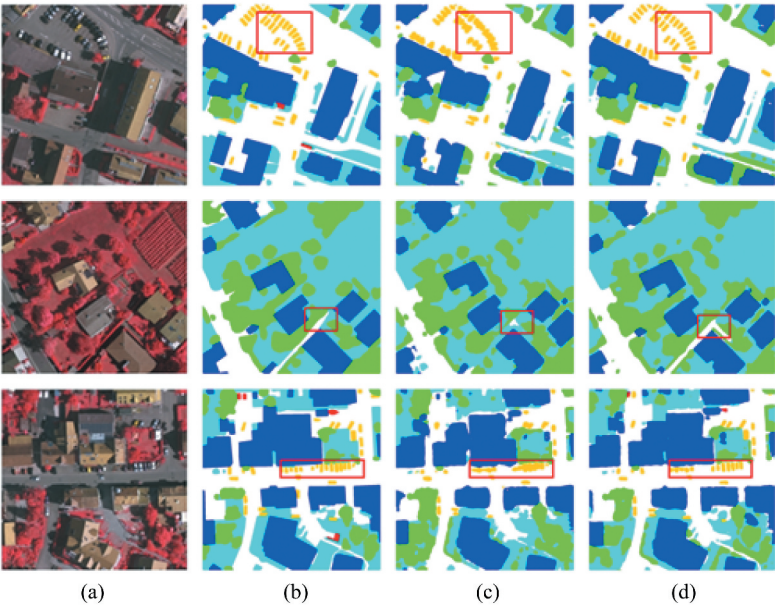


Figure 13. Comparison of segmentation results before and after using LKA-PM in the Swin-Unet framework. (a) Image (b) GT (c) Swin-Unet (d) Swin-Unet + LKA-PM.

Table 4. Ablation experiments of the proposed modules on the Potsdam dataset.

Model	F1(%)					Evaluation index	
	Surface	Building	Low	Tree	Car	MF1	MIoU
Baseline	93.4	96.1	88.3	88.4	95.5	92.1	87.0
Baseline + LKA-PM	93.1	96.6	88.3	88.6	96.3	92.5	87.3
Baseline + Coordswin	94.0	96.2	88.0	88.8	95.8	92.6	86.5
Baseline + WORC	93.6	96.9	88.3	88.5	95.3	92.3	87.2
Baseline + LKA-PM + WORC	94.1	97.6	88.6	89.3	96.0	93.6	87.7
Baseline + Coordswin + WORC	93.9	97.0	88.2	88.7	95.6	93.5	87.0
Baseline + LKA-PM + Coordswin	93.9	97.4	88.3	90.0	96.2	93.1	87.5
Baseline + LKA-PM+ Coordswin+ WORC	94.4	97.4	88.4	89.2	96.8	93.8	87.9

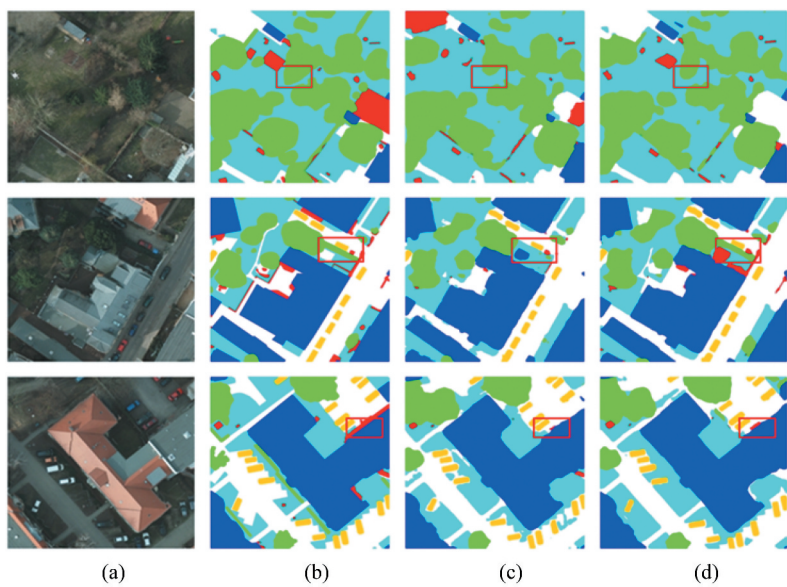


Figure 14. Comparison of segmentation results before and after using Coordswin in the Swin-UNet framework. (a) Image (b) GT (c) Swin-UNet (d) Swin-UNet + Coordswin.

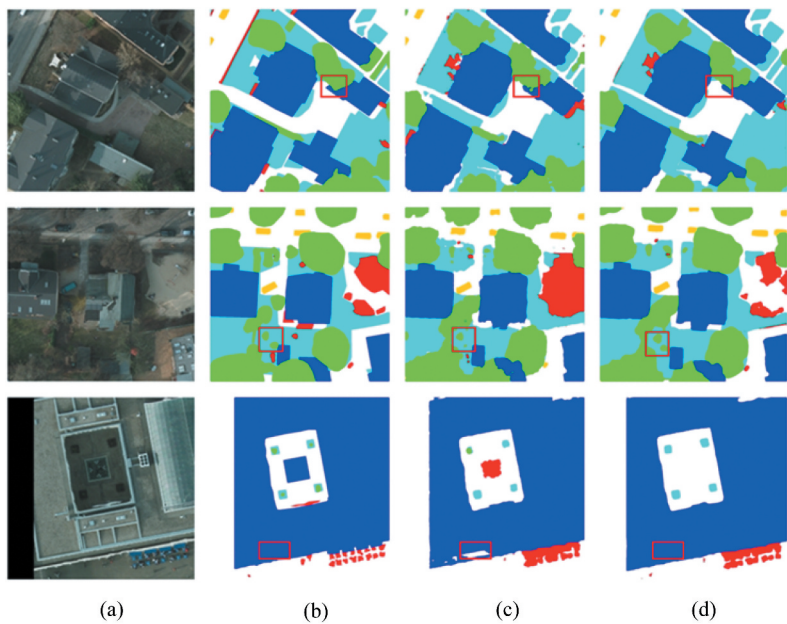


Figure 15. Comparison of segmentation results before and after using WORC in the Swin-UNet framework. (a) Image (b) GT (c) Swin-UNet (d) Swin-UNet + WORC.

Table 5. Comparison of model parameters and accuracy.

Model	Params(MB)	FLOPs(Gbps)	Potsdam		Vaihingen	
			MF1	MIoU	MF1	MIoU
MANet	40.9	65.5	92.9	87.0	90.4	82.7
SwiftNet	35.8	55.6	91.0	83.8	88.3	79.6
DeepLabV3+	42.2	182.2	90.9	84.2	86.7	77.1
SwinUpperNet	102.2	120.6	92.2	85.8	89.8	81.8
FTUNetFormer	91.0	132.5	93.3	87.5	91.3	84.1
DC-Swin	45.6	48.5	92.7	86.5	90.7	83.2
MCAT-UNet	23.5	18.5	92.9	86.8	90.5	83.7
DEDNet	85.5	97.5	93.0	86.9	90.8	83.5
CW-SwinUNet	77.3	89.2	93.8	87.9	91.8	85.2

there is a 0.3 % increase in MIoU and a 0.4 % increase in Ave.F1. Among the categories, the largest improvement in classification accuracy is observed in the ‘low vegetation’ category, with a 0.6 % increase. A visual comparison of the segmentation results is shown in Figure 14.

In the second and third rows, the use of the Coordswin module prevents misclassifications caused by variations in lighting on impermeable surfaces and low-height vegetation. In the first row, where trees and low vegetation are closely positioned, their similarity makes accurate segmentation challenging. However, with the utilization of Coordswin, they can be accurately segmented. Experiments demonstrate that incorporating additional coordinate information through CoordConv can effectively improve the segmentation accuracy of highly similar objects.

Effect of Reparameterization: As shown in Table 3, when WORC is used alone, the model exhibits improvements of 0.4 % in MIoU and 0.9 % in Ave.F1. This validates the effectiveness of WORC in our network. Due to the occlusion of the ‘impervious surface’ class by the ‘tree’ class in remote sensing (RS) imagery, extracting and recognizing the semantic features of the former becomes challenging. Consequently, the model struggles to precisely identify the regions corresponding to ‘trees’.

Figure 15 shows that introducing WORC effectively enhances the model’s ability to capture features across multiple scales and resolutions. Table 3 shows that, on the Potsdam dataset, simultaneous incorporation of LKA-PM, WORC, and Coordswin yields improvements of 2.3 % in mIoU and 1.4 % in Ave.F1. The joint effects of these modules were studied within the CW-SwinUnet framework, as depicted in Table 3. When LKA-PM and WORC were introduced simultaneously, there was an increase of 2.0 % in MIoU and 1.1 % in Ave.F1. When both WORC and Coordswin were included, the segmentation results improved by 1.3 % in MIoU and 0.5 % in Ave.F1. Moreover, considering LKA-PM and Coordswin concurrently led to a 2.1 % increment in MIoU and a 1.2 % increment in Ave.F1. Furthermore, as demonstrated in Table 4, simultaneous introduction of LKA-PM, WORC, and Coordswin in the Vaihingen dataset yielded improvements of 0.9 % in MIoU and 1.7 % in Ave.F1. When LKA-PM and WORC were simultaneously incorporated, there was an increase of 0.7 % in MIoU and 1.5 % in Ave.F1. When both WORC and Coordswin were included, the segmentation results improved by 0.6 % in MIoU and 1.4 % in Ave.F1. Moreover, considering LKA-PM and Coordswin concurrently resulted in a 0.5 % increment in MIoU and a 1.1 % increment in Ave.F1.

4. Complexity analysis

To provide a comprehensive comparison, Table 5 lists the parameter count and computational cost of different models under the same operating environment. The parameter count and computational cost are measured in megabytes (MB) and gigabits per second, respectively. From the table, it can be seen that models incorporating Swin Transformer blocks generally have lower parameter counts compared to those with purely CNN structures. Although CW-SwinUNet does not have the lowest parameter count or computational cost, it achieves the best results across both datasets (91.8 % F1 and 85.2 % MIoU on the Vaihingen dataset and 93.8 % F1 and 87.9 % MIoU on Potsdam datasets). Evidently, CW-SwinUNet significantly enhances the performance of the method with a relatively small parameter count.

5. Limitation and discussion

Despite achieving a favourable balance between performance and efficiency, our method still has certain limitations. A considerable number of misclassifications still exist. As observed in Figures 11, 12 and the boundary lines of the segmentation results are not sufficiently smooth, failing to match the morphology of elongated objects. This may be attributed to the scarcity of elongated object samples in the Potsdam and Vaihingen datasets. However, remote sensing data volumes are massive, and the labelling of samples is labour-intensive and resource-consuming. Furthermore, single-modal data makes it difficult to accurately identify all ground object information, while multi-modal interaction and data augmentation can provide diverse information. Therefore, a promising future research direction is to optimize algorithms and adopt unsupervised methods, enabling the model to adaptively extract effective features and perform accurate ground object segmentation under unsupervised or weakly supervised conditions. Second, efforts should be made to study complex network modelling and integrate multi-modal features to improve the consistent representation of features. Finally, further improvements can be made in lightweight network design. In summary, potential future improvement directions include: 1) Developing unsupervised semantic segmentation networks to enhance model generalization; 2) Implementing multi-modal feature interaction to strengthen feature expression; 3) Adopting knowledge distillation techniques to reduce the number of model parameters while maintaining performance, thereby making the model more lightweight and reducing computational resource consumption.

6. Conclusions

Remote sensing image semantic segmentation plays a crucial role in extracting meaningful information from satellite imagery. To enhance accuracy and effectiveness, we propose a novel framework, CW-SwinUnet, which integrates the capabilities of the Swin Transformer and U-Net. By adapting the transformer architecture to the semantic segmentation domain, the CW-SwinUnet framework effectively captures both global and local contextual information, leading to more precise segmentation results. Specifically, to extract local details, an LKA-PM module is designed to address the issue of local detail loss in Swin Transformer and enhance the segmentation

capability for small objects in the imagery. Furthermore, the WORC module and Coordswin module are introduced in the decoding process to further improve accuracy. The WORC module enhances detail preservation without increasing computational complexity, effectively distinguishing highly similar objects. The Coordswin module focuses on pixel-level spatial feature correlation, preserving more spatial information and reducing boundary blurring caused by occlusion in remote sensing imagery. The CW-SwinUNet framework has been extensively evaluated on the ISPRS Vaihingen and Potsdam datasets. The results demonstrate that in complex geographical environments, particularly in cases involving irregularities and blurred edges, such as streets, vegetation, and roads. Our proposed method effectively identifies edge regions and small target features. It achieves superior segmentation performance. UTF8gbsn

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Natural Science Foundation of Shanghai Municipality [23ZR1446100]; Shanghai Sailing Program [19YF1437200].

References

- Cai, X., Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao. 2024. "Poly Kernel Inception Network for Remote Sensing Detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA, USA), 27706–27716.
- Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation: 15th European ConferenceVI, Munich, Germany: 833–851. September 8–14, 2018.
- Chen, Y., H. Hua, Y. Zhang, and Z. Yang. 2025. "Ae-Unet++: Attention-Enhanced UNet++ for Building Segmentation of Remote Sensing Images." *Engineering Research Express* 7 (2): 025283. <https://doi.org/10.1088/2631-8695/ade033>.
- Chen, Z., Y. Zhou, B. Wang, X. Xu, N. He, S. Jin, and S. Jin. 2022. "Egde-Net: A Building Change Detection Method for High-Resolution Remote Sensing Imagery Based on Edge Guidance and Differential Enhancement." *ISPRS Journal of Photogrammetry & Remote Sensing* 191:203–222. <https://doi.org/10.1016/j.isprsjprs.2022.07.016>.
- Ding, X., Y. Zhang, Y. Ge, S. Zhao, L. Song, X. Yue, and Y. Shan. 2024. "Unireplknet: A Universal Perception Large-Kernel Convnet for Audio Video Point Cloud Time-Series and Image Recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5513–5524.
- Feng, H., Q. Hu, P. Zhao, S. Wang, M. Ai, D. Zheng, and T. Liu. 2025. "FTransDeepLab: Multimodal Fusion Transformer-Based DeepLabV3+ for Remote Sensing Semantic Segmentation." *IEEE Transactions on Geoscience & Remote Sensing* 63:1–18. <https://doi.org/10.1109/TGRS.2025.3553478>.
- Feng, M., C. Yan, Z. Wu, W. Dong, Y. Wang, and A. Mian. 2025. "Hyperrectangle Embedding for Debaised 3D Scene Graph Prediction from RGB Sequences." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 47 (8): 6410–6426. <https://doi.org/10.1109/TPAMI.2025.3560090>.

- Guo, M.-H., C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu. 2023. "Visual Attention Network." *Computational Visual Media* 9 (4): 733–752. <https://doi.org/10.1007/s41095-023-0364-2>.
- He, X., Z. Yong, J. Zhao, D. Zhang, R. Yao, and Y. Xue. 2022. "Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation." *IEEE Transactions on Geoscience & Remote Sensing* 60:4408715 1–1.
- Hu, L., X. Zhou, J. Ruan, and S. Li. 2024. "ASPP±Planet: A Multi-Scale Context Extraction Network for Semantic Segmentation of High-Resolution Remote Sensing Images." *Remote Sensing* 16 (6): 1036. <https://doi.org/10.3390/rs16061036>.
- Jin, T., K. Chen, S. Yamane, and Y. Kuroda. 2022. "M-DenseUNet: Multi Dense Encoder Connected UNet for Biomedical Image Segmentation." *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)* (Osaka, Japan), 919–921.
- Jing, W., W. Zhang, D. Di, C. Li, M. Emam, and A. Mian. 2025. "Hypergraph Biformer for Semantic Segmentation of High-Resolution Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 63:4406915 1–15.
- Lau, K., L. Po, and Y. Rehman. 2023. "Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in CNN." *Expert Systems with Applications* 236:121352. <https://doi.org/10.1016/j.eswa.2023.121352>.
- Li, R., S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson. 2022. "Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–13. <https://doi.org/10.1109/TGRS.2021.3093977>.
- Li, Y., S. Ouyang, and Y. Zhang. 2022. "Combining Deep Learning and Ontology Reasoning for Remote Sensing Image Semantic Segmentation." *Knowledge-Based Systems* 243:108469. <https://doi.org/10.1016/j.knosys.2022.108469>.
- Li, Y., Y. Zhou, Y. Zhang, L. Zhong, J. Wang, and J. C. Dkdfn. 2022. "Domain Knowledge-Guided Deep Collaborative Fusion Network for Multimodal Unitemporal Remote Sensing Land Cover Classification." 2022, 186 *isprsjprs*, 13:170–189. *ISPRS Journal of Photogrammetry & Remote Sensing* 186:170–189. <https://doi.org/10.1016/j.isprsjprs.2022.02.013>.
- Liu, K., M. Feng, W. Zhao, J. Sun, W. Dong, Y. Wang, and A. Mian. 2025. "Pixel-Level Noise Mining for Weakly Supervised Salient Object Detection." *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.
- Liu, T., Y. Liu, C. Zhang, L. Yuan, X. Sui, and Q. Chen. 2024. "Hyperspectral Image Super-Resolution via Dual-Domain Network Based on Hybrid Convolution." *IEEE Transactions on Geoscience & Remote Sensing* 62:1–18. <https://doi.org/10.1109/TGRS.2024.3370107>.
- Ma, C., R. Mu, M. Li, J. He, C. Hua, L. Wang, J. Liu, G. Totis, J. Yang, K. Liu Zhou Y. Zhou J. Deng X. Weng S. 2025. "A Multi-Scale Spatial–Temporal Interaction Fusion Network for Digital Twin-Based Thermal Error Compensation in Precision Machine Tools." *Expert Systems with Applications* 286:127812. <https://doi.org/10.1016/j.eswa.2025.127812>.
- Meng, W., L. Shan, S. Ma, D. Liu, and B. Hu. 2025. "DLNet: A Dual-Level Network with Self-and Cross-Attention for High-Resolution Remote Sensing Segmentation." *Remote Sensing* 17 (7): 1119. <https://doi.org/10.3390/rs17071119>.
- Miao, R., Y. Zhang, Y. Dong, and B. Du. 2024. "Cross-Task Meta-Learning Network with Graph-Enhanced Attention Module for Hyperspectral Change Detection." *IEEE Transactions on Geoscience & Remote Sensing* 62:1–15. <https://doi.org/10.1109/TGRS.2024.3451457>.
- Rottensteiner, F., G. Sohn, M. Gerke, J. Wegner, U. Breitkopf, and J. Jung. 2014. "Results of the ISPRS Benchmark on Urban Object Detection and 3D Building Reconstruction." *ISPRS Journal of Photogrammetry & Remote Sensing* 93:256–271. <https://doi.org/10.1016/j.isprsjprs.2013.10.004>.
- Sha, X., Z. Guan, Y. Wang, J. Han, Y. Wang, and Z. Chen. 2025. "SSC-Net: A Multi-Task Joint Learning Network for Tongue Image Segmentation and Multi-Label Classification." *Digital Health* 11:20552076251343696. <https://doi.org/10.1177/20552076251343696>.
- Sha, X., X. Si, Y. Zhu, S. Wang, and Y. Zhao. 2025. "Automatic Three-Dimensional Reconstruction of Transparent Objects with Multiple Optimization Strategies under Limited Constraints." *Image and Vision Computing*: 105580. <https://doi.org/10.1016/j.imavis.2025.105580>.

- Tian, Q., F. Zhao, Z. Zhang, and H. Qu. 2023. "GLFFNet: A Global and Local Features Fusion Network with Biencoder for Remote Sensing Image Segmentation." *Applied Sciences* 13:8725. <https://doi.org/10.3390/app13158725>.
- Wang, H., X. Jiang, H. Ren, Y. Hu, and S. Bai. 2021. "Swiftnet: Real-time video object segmentation." 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1296–1305.
- Wang, J., M. Wang, K. Cong, and Z. Qin. 2024. "A Semantic Segmentation Method for Remote Sensing Images Based on an Improved TransDeepLab Model." *The Land* 14 (1): 22. <https://doi.org/10.3390/land14010022>.
- Wang, L., R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang. 2022. "A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images." *IEEE Geoscience & Remote Sensing Letters* 19:1–5. <https://doi.org/10.1109/LGRS.2022.3143368>.
- Wang, L., R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. Atkinson. 2022. "Unetformer: A UNet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery." *ISPRS Journal of Photogrammetry & Remote Sensing* 190:196–214. <https://doi.org/10.1016/j.isprsjsprs.2022.06.008>.
- Wang, T., G. Chen, X. Zhang, C. Liu, J. Wang, X. Tan, W. Zhou, and C. He. 2025. "LMFNet: Lightweight Multimodal Fusion Network for High-Resolution Remote Sensing Image Segmentation." *Pattern Recognition* 164:111579. <https://doi.org/10.1016/j.patcog.2025.111579>.
- Wang, T., C. Xu, B. Liu, G. Yang, E. Zhang, D. Niu, and H. Zhang. 2024. "MCAT-UNet: Convolutional and Cross-Shaped Window Attention Enhanced UNet for Efficient High-Resolution Remote Sensing Image Segmentation." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 17:9745–9758. <https://doi.org/10.1109/JSTARS.2024.3397488>.
- Wang, X., H. Wang, Y. Jing, X. Yang, and J. Chu. 2024. "A Bio-Inspired Visual Perception Transformer for Cross-Domain Semantic Segmentation of High-Resolution Remote Sensing Images." *Remote Sensing* 16 (9): 1514. <https://doi.org/10.3390/rs16091514>.
- Wang, Z., J. Li, Z. Tan, X. Liu, and M. Li. 2023. "Swin-Upernet: A Semantic Segmentation Model for Mangroves and Spartina Alterniflora Loisel Based on Upernet." *Electronics* 12 (5): 1111. <https://doi.org/10.3390/electronics12051111>.
- Wang, Z., M. Xia, L. Weng, K. Hu, and H. Lin. 2024. "Dual Encoder–Decoder Network for Land Cover Segmentation of Remote Sensing Image." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 17:2372–2385. <https://doi.org/10.1109/JSTARS.2023.3347595>.
- Xia, C., X. Wang, F. Lv, X. Hao, and Y. Shi. 2024. "Vit-Comer: Vision Transformer with Convolutional Multi-Scale Feature Interaction for Dense Predictions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (Seattle, WA, USA)*, 5493–5502.
- Xiang, D., D. He, H. Sun, P. Gao, J. Zhang, and J. Ling. 2025. "HCMPE-Net: An Unsupervised Network for Underwater Image Restoration with Multi-Parameter Estimation Based on Homology Constraint." *Optics and Laser Technology* 186:112616. <https://doi.org/10.1016/j.optlastec.2025.112616>.
- Xu, J., Z. Xiong, and S. P. Bhattacharyya. 2023. "Pidnet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers." 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC, Canada), 19529–19539.
- Zeng, Q., J. Zhou, J. Tao, L. Chen, X. Niu, and Y. Zhang. 2024. "Multiscale Global Context Network for Semantic Segmentation of High-Resolution Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 62:1–13. <https://doi.org/10.1109/TGRS.2024.3393489>.
- Zhang, H., Z. Jiang, G. Zheng, and X. Yao. 2023. "Semantic Segmentation of High-Resolution Remote Sensing Images With Improved U-Net Based on Transfer Learning." *International Journal of Computational Intelligence Systems* 16 (1): 181. <https://doi.org/10.1007/s44196-023-00364-w>.
- Zhang, R., Q. Zhang, and G. Zhang. 2024. "Lsrformer: Efficient Transformer Supply Convolutional Neural Networks with Global Information for Aerial Image Segmentation." *IEEE Transactions on Geoscience & Remote Sensing* 62:5610713 1–13.
- Zhu, E., Z. Chen, D. Wang, H. Shi, X. Liu, and L. Wang. 2025. "Unetmamba: An Efficient UNet-Like Mamba for Semantic Segmentation of High-Resolution Remote Sensing Images." *IEEE Geoscience & Remote Sensing Letters* 22:1–5. <https://doi.org/10.1109/LGRS.2024.3505193>.

- Zhu, G., and S. C. Kim. 2021. "Coord-Fcn for Same-Class Objects Segmentation." 2021 International Conference on Information and Communication Technology Convergence (ICTC) (Jeju Island, South Korea), 1672–1674.
- Zhu, S., L. Zhao, Q. Xiao, J. Ding, and X. Li. 2025. "GLFFNet: Global–Local Feature Fusion Network for High-Resolution Remote Sensing Image Semantic Segmentation." *Remote Sensing* 17 (6): 1019. <https://doi.org/10.3390/rs17061019>.
- Zhu, X., J. Lu, H. Ren, H. Wang, and B. Sun. 2022. "A Transformer–CNN for Deep Image Inpainting Forensics." *The Visual Computer* 39 (10): 39. <https://doi.org/10.1007/s00371-022-02620-0>.