Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 33(10): 6391–6417.
DOI: 10.3934/era.2025282
Received: 12 September 2025
Revised: 20 October 2025
Accepted: 21 October 2025
Published: 28 October 2025

Research article

Spatial structure-aware and cross-scale feature modeling network for remote sensing image semantic segmentation

Fangbin Huang*and Yuxuan Guo

School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China

* **Correspondence:** Email: 202212490416@nuist.edu.cn.

Abstract: Remote sensing images exhibit significant spatial geometric characteristics for ground objects such as buildings and roads, while targets within scenes show enormous scale variations, posing challenges to semantic segmentation algorithms' spatial structure modeling capabilities and cross-scale information processing abilities. Traditional methods lack specialized modeling mechanisms for spatial geometric features and suffer from information loss in multi-scale feature fusion. This paper proposes the SC-Net network, addressing these issues through three key technological innovations. First, we designed a feature attention layer where the spatial attention module captures spatial geometric patterns through directional feature decomposition, and the multi-scale attention module preserves feature information at different scales through adaptive pooling strategies. Second, we constructed a three-branch fusion transformer that employs cross-window attention and nine-group feature key-value pair interactions to achieve collaborative modeling of spatial, multi-scale, and global features. Finally, the multi-branch cascaded decoder enhances segmentation boundary accuracy through hierarchical feature fusion strategies. Comprehensive experiments on three standard remote sensing datasets validated the method's superiority. SC-Net achieved 63.04% mean intersection over union (MIOU) on Wuhan dense labeling dataset (WHDLD), 71.57% on Potsdam dataset, and 81.57% on Vaihingen dataset, outperforming state-of-the-art methods such as AerialFormer and SERNet by 0.67–2.12% MIOU. The method particularly demonstrated outstanding performance in scenarios with complex spatial structures and dense multi-scale targets, providing an effective solution for precise remote sensing image interpretation.

Keywords: remote sensing images; semantic segmentation; cross-scale feature fusion; spatial structure modeling; transformer; attention mechanism

1. Introduction

The rapid development of remote sensing technology has made high-quality remote sensing images play an increasingly important role in geological detection, land classification, and environmental monitoring [1]. Remote sensing image semantic segmentation, as a key technology for understanding ground object scenes at the pixel level, aims to assign accurate semantic category labels to each pixel, thereby realizing effective extraction and utilization of rich spatial information in remote sensing images [2].

Remote sensing image semantic segmentation faces two interrelated core challenges, as shown in Figure 1. First is the problem of insufficient spatial structure information modeling—buildings typically present regular rectangular or polygonal structures, roads exhibit continuous linear features, while vegetation areas display relatively irregular areal distributions [3, 4]. However, existing methods often treat these targets with essentially different spatial geometric characteristics equally, lacking targeted spatial structure modeling capabilities [5]. Meanwhile, the problems of cross-scale information loss and difficulty in detail preservation are also becoming increasingly prominent—due to the long shooting distance and wide field of view of remote sensing images, targets in the same scene exhibit significant scale variations. Large buildings may occupy hundreds of pixels, while small vehicles or trees may only have a few pixels in size [6]. During the downsampling process of deep networks, feature information of small-scale targets is extremely prone to loss, and traditional multi-scale processing methods are often passive and coarse-grained, unable to adaptively adjust feature extraction strategies based on image content [7]. These two challenges jointly lead to inaccurate boundary localization when the model processes targets with obvious geometric features, insufficient recognition capability for small targets, and limited segmentation accuracy in complex scenes. Figure 1(a) demonstrates the ambiguity of pixel features in category boundary regions, Figure 1(b) reflects the difficulty of multi-scale target recognition, and Figure 1(c) highlights the modeling challenges of targets with different spatial characteristics.

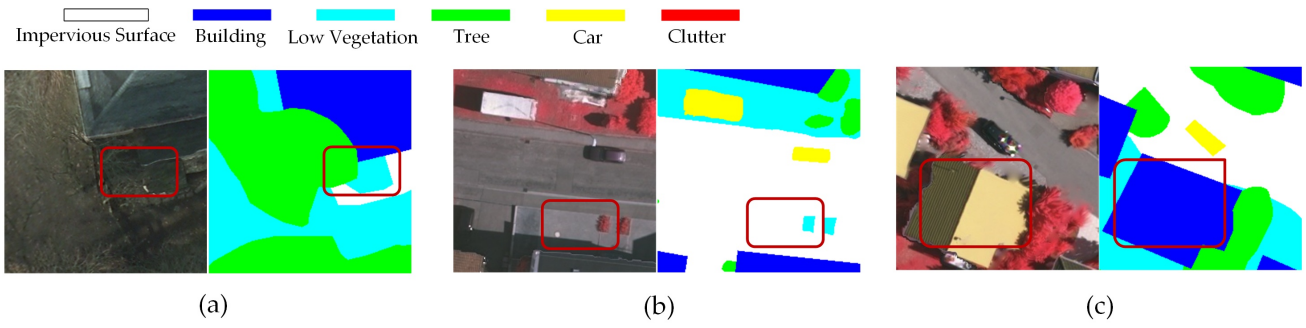

Figure 1. Illustration of the main challenges in remote sensing image semantic segmentation.

The rise of deep learning has brought new development opportunities for remote sensing image semantic segmentation. Fully convolutional networks (FCNs) [8] pioneered end-to-end pixel-level classification, while U-Net [9] and DeepLabV3+ [10] achieved significant progress in feature preservation and multi-scale modeling through techniques such as skip connections and atrous convolutions. In recent years, attention mechanism-based methods have been widely applied, such as enhancing feature expression capabilities through spatial attention and channel attention [11], and designing multi-scale feature fusion modules to handle scale variation problems [12].

The emergence of transformer architectures has opened new pathways for remote sensing image

analysis [13]. Its core self-attention mechanism can establish long-range dependencies and effectively capture global contextual information [14]. Segmentation transformers [15] first applied the pure transformer to semantic segmentation, while the SegFormer [16] improved segmentation accuracy through hierarchical feature fusion. However, the full-image self-attention mechanism of standard transformers has excessively high computational complexity, limiting their application on large-scale remote sensing images [17]. To address this issue, the Swin transformer [18] proposed a hierarchical attention mechanism based on sliding windows, and the CSWin Transformer [19] further improved modeling efficiency through cross-stripe window design.

The high-resolution network (HRNet) [20] demonstrates unique advantages in maintaining high-resolution representations through multi-branch parallel architecture and has been widely applied in remote sensing image segmentation [21, 22]. Some improved methods integrate attention mechanisms into the HRNet to enhance feature expression [23, 24], while others explore hybrid CNN-transformer architectures to balance local details and global modeling [25]. However, existing CNN-transformer fusion methods mainly rely on simple feature concatenation or weighted fusion strategies [26], lacking in-depth modeling of complementary relationships between different feature types, particularly still having deficiencies in collaborative expression of spatial structure information and multi-scale features in remote sensing images [27, 28].

To address the above challenges, this paper proposes the SC-Net network, with three key technological innovations specifically targeting the identified problems:

1). Design of the feature attention layer (FAL): To address the issue of existing methods uniformly processing objects with different geometric features, we propose spatial attention (SA) and multi-scale attention (MA) mechanisms. SA enhances attention to spatial feature categories such as buildings through feature slicing and hierarchical processing, while MA reduces small-scale feature information loss through pooling operations at different scales and learnable feature storage units. Unlike existing attention mechanisms (CBAM [29], DAN [30]), our design is specifically tailored for the spatial geometric characteristics of remote sensing images.

2). Proposal of the three-branch fusion transformer (TFT) architecture: To address the limitation that existing CNN-transformer fusion methods rely on simple concatenation strategies, we design an innovative nine-group feature key-value pair interaction mechanism based on the cross-rectangular window partitioning strategy of the CSWin transformer, achieving efficient fusion of spatial, multi-scale, and global features while reducing computational complexity and establishing explicit correlation relationships among heterogeneous feature types.

3). Construction of a multi-branch cascaded decoder: To address the detail loss problem caused by traditional single-path upsampling, we adopt a progressive feature reconstruction strategy from low resolution to high resolution, obtaining more accurate category prediction results and boundary localization through layer-by-layer transmission of feature information between branches.

2. Related work

2.1. Deep learning and transformer methods for remote sensing image semantic segmentation

Deep learning has fundamentally transformed remote sensing image semantic segmentation, evolving from traditional machine learning approaches to sophisticated neural architectures capable of extracting rich hierarchical features [31]. Recent developments have focused on addressing specific challenges

in remote sensing imagery. Wang et al. proposed AFF-UNet with adaptive feature fusion modules to handle the diverse scales and complex backgrounds typical in aerial imagery [32]. Li et al. developed MFCA-Net, which combines multi-feature fusion with channel attention mechanisms to improve small object recognition [33]. These methods demonstrate the ongoing evolution toward more specialized architectures that account for the unique characteristics of remote sensing data, including varying illumination conditions, complex spatial arrangements, and significant scale variations [34, 35].

The integration of attention mechanisms has become increasingly important in recent remote sensing segmentation approaches. Methods such as SCAttNet and SAPNet have shown that spatial and channel attention can significantly improve feature representation by focusing computational resources on the most informative regions [36, 37]. However, most existing CNN-based methods still face limitations in capturing long-range spatial dependencies and global contextual relationships, which has motivated the exploration of transformer-based architectures [2].

The introduction of vision transformers (ViTs) has marked a paradigm shift in remote sensing image analysis, offering superior capability in modeling long-range dependencies and global contextual relationships [38]. Unlike CNNs that process images through localized convolution operations, transformers utilize self-attention mechanisms to establish relationships between all spatial positions simultaneously, making them particularly suitable for remote sensing images where global context is often crucial for accurate interpretation [39].

The computational challenges of standard transformers have led to several specialized architectures for remote sensing applications. These hierarchical attention mechanisms have proven particularly effective for remote sensing tasks where objects appear at multiple scales and orientations [14]. Recent work has explored hybrid CNN-transformer architectures that combine the advantages of both paradigms. UNetFormer integrates transformer blocks within U-Net structures to capture both local spatial details and global contextual information [40]. CTFuseNet demonstrates the effectiveness of parallel CNN and transformer branches for crop type segmentation, showing improved performance over single-architecture approaches [25]. However, most existing fusion strategies rely on simple concatenation or weighted averaging, potentially limiting the full utilization of complementary information from different architectural components.

Recent advances in few-shot semantic segmentation have explored learning with limited labeled samples through innovative attention mechanisms. For instance, class-aware self- and cross-attention approaches [41] have demonstrated effective strategies for capturing robust class information from support samples in remote sensing scenarios, while global-local query-support frameworks [42] have shown how to leverage both prototype-level semantic information and pixel-level local details for improved segmentation. While our work addresses the standard fully-supervised scenario with sufficient training data, these few-shot approaches provide valuable insights into attention mechanism design and feature interaction strategies that could inform future extensions toward data-efficient remote sensing segmentation.

2.2. Multi-scale feature fusion and attention mechanisms

Multi-scale feature fusion has emerged as a critical component in remote sensing image analysis due to the significant scale variations inherent in aerial and satellite imagery [43]. Traditional approaches to multi-scale processing, such as image pyramids and multi-resolution analysis, have evolved into sophisticated deep learning architectures that can adaptively extract and combine features across different scales [44].

Feature pyramid networks (FPNs) established the foundation for modern multi-scale fusion by creating a top-down pathway with lateral connections to combine high-level semantic information with low-level spatial details [45]. This architecture has been widely adopted in remote sensing applications, with numerous improvements proposed to address specific challenges. The multi-scale feature progressive fusion network (MFPF-Net) introduced layer feature fusion modules and multi-scale feature aggregation to improve change detection accuracy. Similarly, MS2LandsNet demonstrated the effectiveness of lightweight multi-scale attention mechanisms for efficient landslide detection in medium-resolution imagery [46].

Attention mechanisms have been increasingly integrated with multi-scale architectures to improve feature selection and fusion. The CBAM and its variants provide both channel and spatial attention to enhance feature representation. However, these methods typically operate on individual feature maps without considering the relationships between different scales. More recent work has explored hierarchical attention mechanisms that can model dependencies across multiple resolution levels. The multi-scale channel and spatial attention module (MCAM) demonstrates improved building extraction by jointly considering multi-scale spatial and channel relationships [47].

Advanced fusion strategies have moved beyond simple concatenation or addition operations toward more sophisticated integration mechanisms. The hierarchical multi-scale feature fusion network (MHF2Net) introduces self-weighted attention blocks that automatically determine the contribution of different scale features without requiring supervised parameters [48]. Multi-scale dense graph attention networks have shown promise for hyperspectral image classification by modeling complex feature relationships across different spectral and spatial scales.

Despite these advances, most existing multi-scale fusion methods face several limitations. First, they often treat different feature types uniformly without considering their unique characteristics and complementary relationships [49]. Second, the fusion process typically occurs at fixed network layers, potentially missing optimal integration points for different semantic concepts [50]. Third, current attention mechanisms primarily focus on enhancing individual features rather than modeling the interactions between different feature types and scales [5]. Recent work has begun to address these limitations through more sophisticated fusion architectures, with adaptive fusion strategies that can dynamically adjust the contribution of different features based on input content showing promise for handling diverse remote sensing scenarios [26, 49]. However, the development of unified frameworks that can effectively integrate spatial structural information, multi-scale features, and global contextual relationships remains an active area of research [50].

Attention mechanisms have been increasingly integrated with multi-scale architectures to improve feature selection and fusion. The CBAM and its variants provide both channel and spatial attention to enhance feature representation, while methods like efficient channel attention (ECA) [51] and coordinate attention (CA) [52] have explored efficient designs for modeling inter-channel and positional relationships. However, these methods are primarily developed for natural images and typically operate on individual feature maps without explicitly considering the unique characteristics of remote sensing imagery. Remote sensing scenes present distinct challenges: buildings exhibit regular rectangular structures, roads demonstrate continuous linear features, and objects show extreme scale variations due to aerial perspective—small vehicles may occupy merely a few pixels while large buildings span hundreds of pixels. Existing attention mechanisms lack specialized modeling for these directional geometric patterns and struggle to preserve small-scale features during downsampling. Our feature

attention layer addresses these challenges through targeted design, where the spatial attention module captures directional geometric features through separate X-axis and Y-axis processing, while the multi-scale attention module employs adaptive pooling at multiple scales with learnable feature storage to prevent information loss of small objects during network depth progression.

3. Methodology

3.1. Overall network architecture

This section details SC-Net's overall architecture and its key module design principles. To address the two core challenges of insufficient spatial structure modeling and cross-scale information loss in remote sensing image semantic segmentation, we propose a complete solution that enhances the expression of spatial geometric features and multi-scale features through feature attention layers, achieves efficient fusion of heterogeneous features through a three-branch fusion transformer, and implements progressive feature reconstruction through a multi-branch cascaded decoder.

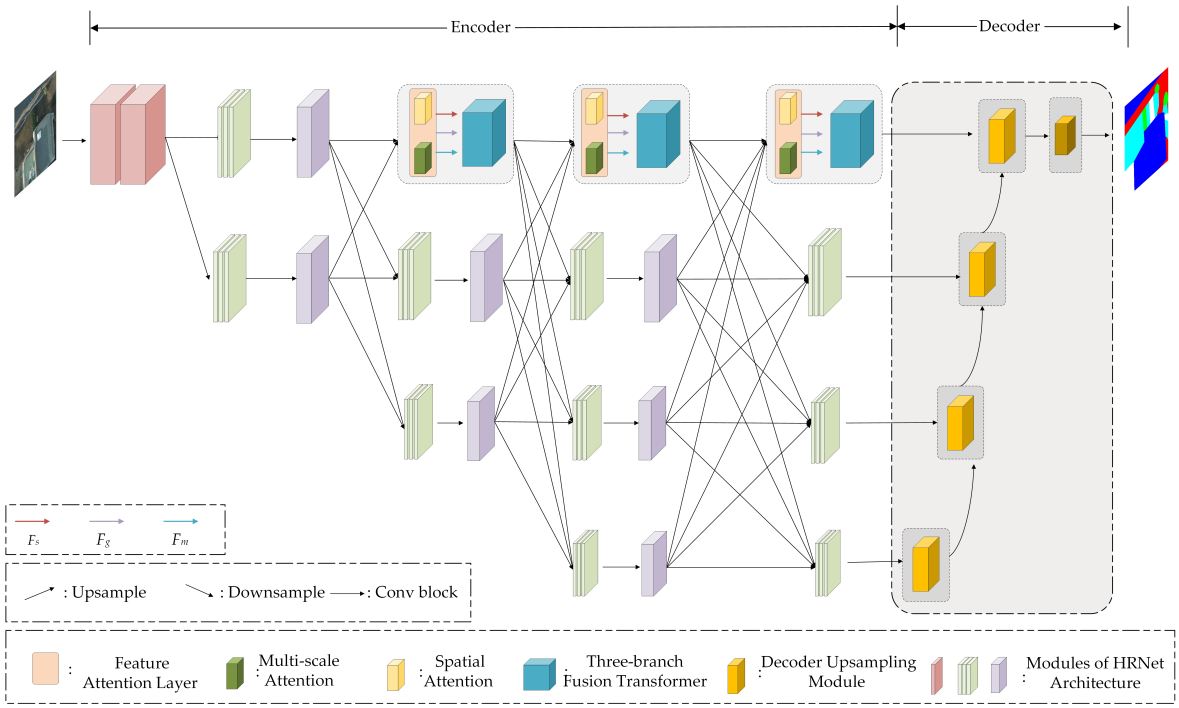

Figure 2. Overall architecture of SC-Net showing the multi-branch encoder, feature attention layers (FAL), three-branch fusion transformer (TFT), and multi-branch cascaded decoder.

Figure 2 illustrates the overall architecture of SC-Net. The network employs a multi-branch parallel encoder based on HRNet as the backbone network, which can simultaneously extract and maintain multi-scale feature representations across different resolution branches. Unlike traditional methods that primarily focus on multi-scale feature fusion, our design emphasizes modeling the inherent spatial structures and semantic relationships in remote sensing images.

The encoder consists of four parallel branches corresponding to different resolution levels, achieving collaborative learning of multi-scale features through information exchange between branches. After

each HRNet module, we introduce feature attention layers (FAL) to specifically handle the enhancement of spatial geometric features and multi-scale features. The output of the FAL serves as input to the three-branch fusion transformer (TFT), which achieves deep fusion of spatial features, multi-scale features, and global features through innovative cross-rectangular window mechanisms and nine-group key-value pair interactions.

Unlike the simple upsampling approach of the original HRNet, we design a multi-branch cascaded decoder to achieve more precise feature decoding. The decoder adopts a progressive fusion strategy, starting from low-resolution branches and gradually transferring and refining feature information to high-resolution branches, ensuring effective recovery of detailed information.

3.2. Feature attention layer

Different ground object categories in remote sensing images exhibit distinct spatial distribution characteristics and scale variations. Traditional methods often treat these heterogeneous features uniformly, leading to loss of spatial structural information and fine-scale information. To address this problem, we design the feature attention layer (FAL), which specifically enhances spatial geometric features and multi-scale features through two complementary modules: spatial attention (SA) and multi-scale attention (MA).

3.2.1. Spatial attention mechanism

Objects such as buildings and roads in remote sensing images exhibit obvious spatial geometric regularity, and this spatial structural information is crucial for accurate segmentation. Traditional attention mechanisms mainly focus on channel or global spatial relationships, lacking the ability to model specific spatial directions and geometric patterns.

As shown in Figure 3, spatial attention (SA) employs a parallel branch structure to extract latent spatial features. We treat the input features as three-dimensional spatial representations and divide them into g slices along the channel dimension, with each slice independently processed in spatial dimensions. This design enables the model to focus specifically on spatial patterns in different directions, particularly the rectangular structures of buildings and linear features of roads.

For the g-th feature slice, SA first applies SoftPool operations in both the X-axis and Y-axis directions to capture directional features, and then performs feature encoding through 1×1 convolution and sigmoid activation functions. Meanwhile, we apply 1×1 convolution to the original features to concentrate channel dimension information. Subsequently, X-axis, Y-axis, and C-axis dimension features are encoded through Softmax and average pooling operations to generate more refined spatial representations.

Mathematically, this process can be expressed as:

$$X_g = \text{Sigmoid}(\text{Conv}_{1\times1}(\text{SoftPool}_X(F_g))) \odot F_g \tag{3.1}$$

$$Y_g = \text{Sigmoid}(\text{Conv}_{1\times1}(\text{SoftPool}_Y(F_g))) \odot F_g \tag{3.2}$$

$$C_g = \text{Conv}_{1\times1}(F_g) \tag{3.3}$$

$$SA_g = X_g + Y_g + C_g \tag{3.4}$$

where F_g and SA_g represent the input and output of the g-th feature slice, respectively. Finally, we

concatenate all group slices as the output of SA:

$$SA_{out} = \text{Concat}([SA_1, SA_2, \ldots, SA_g]) \tag{3.5}$$

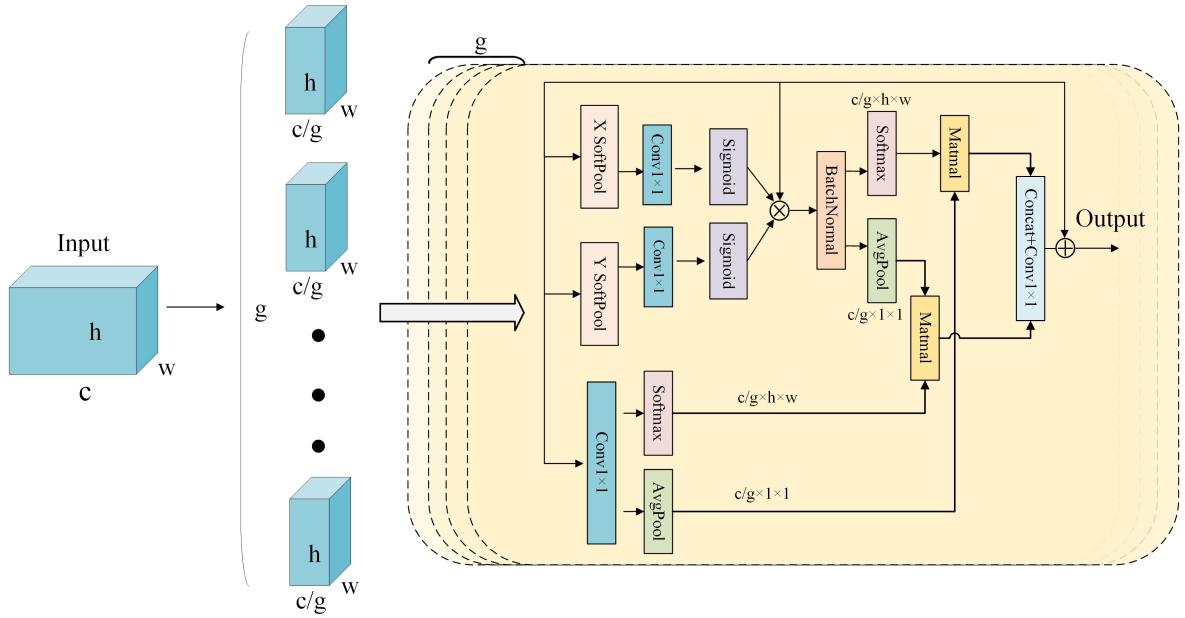

Figure 3. Spatial attention (SA) mechanism with parallel branch structure for extracting directional spatial features.

3.2.2. Multi-scale attention mechanism

The long imaging distance and wide coverage of remote sensing images result in significant scale differences among objects in the same scene. Traditional downsampling operations easily cause loss of small-scale object features, affecting segmentation accuracy. Multi-scale attention (MA) adaptively addresses this problem by designing pooling modules at different scales.

As shown in Figure 4, MA performs downsampling at four different scales on input features, generating features of sizes 1, 2, 4, and 8, and then concatenates them to obtain multi-scale features of the image. To better learn and encode multi-scale features, we introduce learnable feature storage units D, which can store learned main feature information to accelerate model convergence.

This process can be expressed as:

$$M_n = \text{AvgPool}_n(F_{backbone}) \tag{3.6}$$

$$M_{concat} = \text{Concat}([M_1, M_2, M_4, M_8]) \tag{3.7}$$

$$MA_{out} = D \odot M_{concat} + M_{concat} \tag{3.8}$$

where $F_{backbone}$ is the backbone feature extracted from the high-resolution branch of SC-Net, n represents the size after downsampling, and D is the learnable spatial feature storage unit.

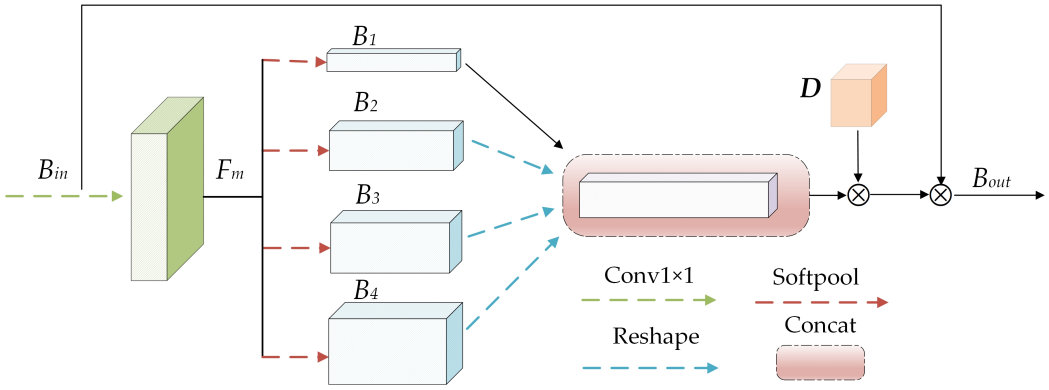

Figure 4. Multi-scale attention (MA) mechanism with different scale pooling operations and learnable feature storage units.

3.3. Three-branch fusion transformer

Traditional feature fusion methods mainly rely on simple feature concatenation or weighted summation, which limits the full utilization of complementary relationships between different feature types. To achieve deep fusion of spatial features, multi-scale features, and global features, we propose the three-branch fusion transformer (TFT).

The TFT combines innovative cross-rectangular window partitioning strategies with complex nine-group feature key-value pair interaction mechanisms, not only reducing computational complexity but more importantly establishing explicit correlation relationships between heterogeneous feature types, significantly enhancing the model's representational capability.

3.3.1. Window partitioning strategy

Standard transformers perform self-attention computation on entire images, leading to excessively high computational complexity (Figure 5(a)). The Swin transformer reduces complexity through sliding square windows but fixed window sizes limit global modeling capability (Figure 5(b)). The CSWin transformer adopts cross-stripe window design (Figure 5(c)), achieving better global feature representation through point-center feature encoding.

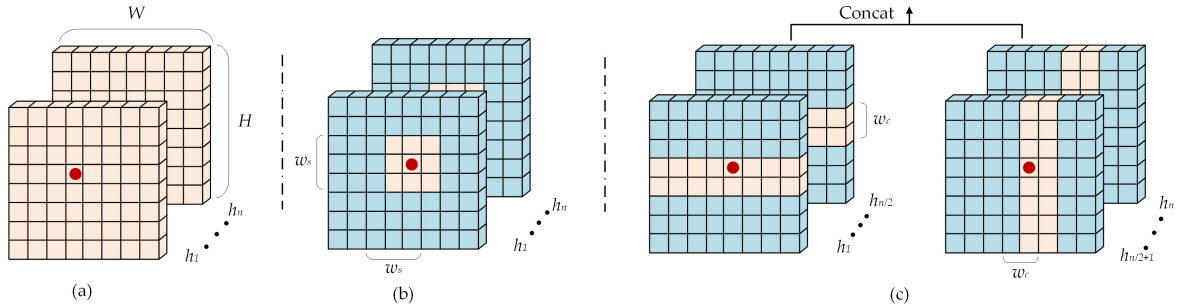

Figure 5. Comparison of attention mechanisms: (a) standard transformer, (b) Swin transformer, (c) CSWin transformer with cross-shaped windows.

Our TFT inherits the advantages of the CSWin transformer's cross-rectangular window partitioning but innovatively designs a three-branch feature fusion mechanism. Unlike standard methods where each

attention head processes the same type of features, we assign three different features from MA, SA, and their matrix sum to different branches for processing.

3.3.2. Nine-group key-value pair interaction mechanism

As shown in Figure 6, The TFT processes multi-scale features, spatial features, and global features through three branches respectively. The input of each branch generates a corresponding Query, Key, and Value through linear transformations:

- Multi-scale feature branch: $F_m \rightarrow (q_m, k_m, v_m)$
- Spatial feature branch: $F_s \rightarrow (q_s, k_s, v_s)$
- Global feature branch: $F_g \rightarrow (q_g, k_g, v_g)$

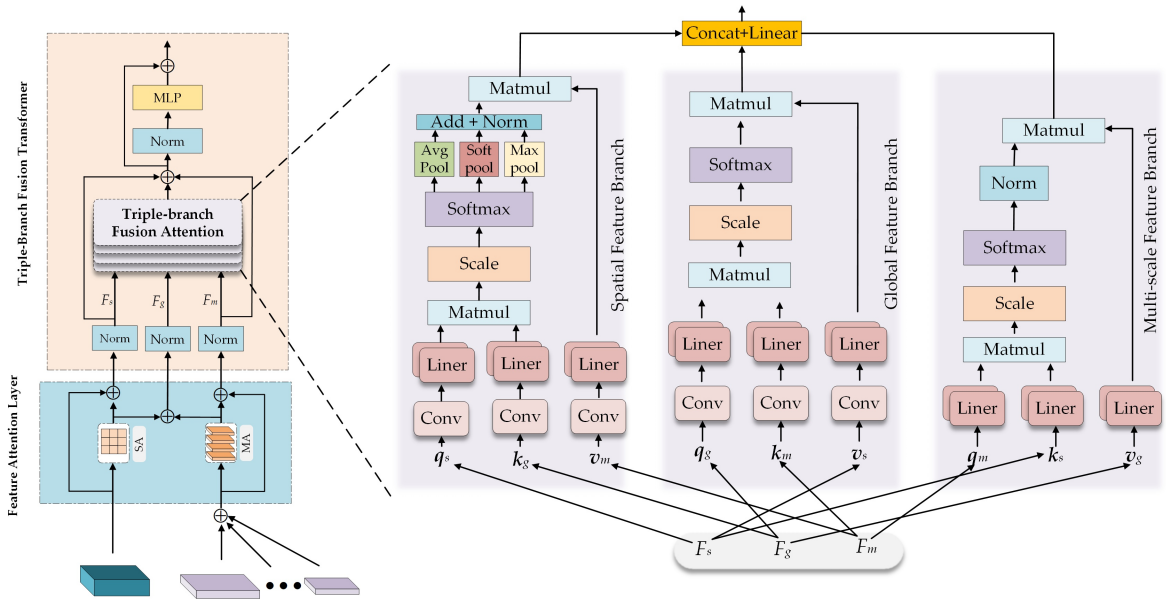

Figure 6. Three-branch fusion transformer (TFT) architecture with a nine-group key-value pair interaction mechanism.

The core innovation of the TFT lies in the cross-use of key-value pairs from different branches. For the multi-scale feature branch, we use the combination of q_m, k_s, v_g; for the spatial feature branch, we use the combination of q_s, k_g, v_m; and for the global feature branch, we use the combination of q_g, k_m, v_s. This cross-configuration ensures deep interaction between different feature types.

For the i-th attention head in the multi-scale feature branch, feature keys and values are divided into different cross-windows through linear transformation:

$$[q_{im}, k_{is}, v_{ig}] = \text{Linear}([q_m, k_s, v_g]) \tag{3.9}$$

where the number of windows is $N = \frac{H \times w_c}{w_c \times H}$, and w_c and H are the width and length of the cross-windows, respectively.

Multi-scale information is obtained through feature encoding:

$$\text{Attention}_{im} = \text{Softmax}\left(\frac{q_{im}k_{is}^T}{\sqrt{d_k}}\right)v_{ig} \tag{3.10}$$

$$M = \text{Concat}([\text{Att}_{1m}, \text{Att}_{2m}, \ldots, \text{Att}_{Im}]) \tag{3.11}$$

Similarly, the spatial feature branch and global feature branch compute, respectively:

$$\text{Attention}_{is} = \text{Softmax}\left(\frac{q_{is}k_{ig}^T}{\sqrt{d_k}}\right)v_{im} \tag{3.12}$$

$$S = \text{Concat}([\text{Att}_{1s}, \text{Att}_{2s}, \ldots, \text{Att}_{Is}]) \tag{3.13}$$

$$\text{Att}_{ig} = \text{Softmax}\left(\frac{q_{ig}k_{im}^T}{\sqrt{d_k}}\right)v_{is} \tag{3.14}$$

$$G = \text{Concat}([\text{Att}_{1g}, \text{Att}_{2g}, \ldots, \text{Att}_{Ig}]) \tag{3.15}$$

Finally, the outputs of the three branches are concatenated along the window number dimension, and then adapted to $C \times H \times W$ dimensions through linear embedding and reshaping operations, maintaining consistency with original feature dimensions.

To ensure reproducibility, we provide complete specifications of the TFT implementation. The attention head dimension at each stage is computed as $d_k = C/h$, where C represents the channel dimension from Table 2 (c_1, c_2, c_3, c_4 for the four stages) and h denotes the number of attention heads ($h_1 = 4, h_2 = 4, h_3 = 8$ for SC-L as shown in Table 2). We employ layer normalization in a pre-norm configuration, applying it before the attention operation and before the feed-forward network. Relative position bias is used for positional encoding, learning a bias term for each relative position within the cross-shaped windows, similar to the approach in the Swin transformer. The feed-forward network (MLP) consists of two linear layers with GELU activation, where the hidden dimension is expanded by a factor of 4 (i.e., intermediate dimension = $4C$). Residual connections are applied around both the attention module and the MLP block. The complete computation for each branch follows: $\mathbf{X}' = \mathbf{X} + \text{Attention}(\text{LN}(\mathbf{X}))$, and $\mathbf{X}'' = \mathbf{X}' + \text{MLP}(\text{LN}(\mathbf{X}'))$, where LN denotes layer normalization. The dropout rate is set to 0.1 for both attention weights and the MLP output. All three branches share this identical architectural structure, with the distinguishing feature being the cross-branch key-value pair assignments described in Eqs (3.9)–(3.15).

3.4. Multi-branch cascaded decoder

Remote sensing image semantic segmentation requires precise boundary localization and fine-grained detail preservation. Traditional single upsampling methods often lead to a loss of spatial details. We design an advanced multi-branch cascaded decoder architecture that starts feature reconstruction from low-resolution branches and systematically propagates refined feature information to higher-resolution branches, achieving hierarchical feature enhancement.

As shown in the decoder part of Figure 2, our method sequentially processes low-resolution features from the encoder and integrates them into preceding high-resolution branches to ensure robustness. Each decoder module implements batch normalization to optimize feature distribution uniformity, followed by 1×1 convolution for channel dimension reduction and bilinear interpolation for $2\times$ upsampling.

This progressive decoding strategy can fully utilize complementary information from different resolution levels, gradually refining from coarse-grained semantic information to fine-grained spatial details, ultimately achieving high-quality segmentation results.

3.5. Loss function

To optimize network training, we adopt a combination of cross-entropy loss and Dice loss, both widely applied in semantic segmentation tasks. Cross-entropy loss quantifies pixel-level deviations from ground truth label values, while Dice loss measures category-level distribution differences from target distributions. Our composite loss function is defined as:

$$L_{total} = L_{CrossEntropy} + L_{Dice} \tag{3.16}$$

This combination of loss functions can simultaneously consider pixel-level accuracy and region-level consistency, facilitating better segmentation performance.

3.6. Architecture variants and module parameters

To validate the effectiveness of the proposed SC-Net architecture, we construct three architecture variants by adjusting specific module parameters in the model: SC-net-tiny (SC-T), SC-net-small (SC-S), and SC-net-large (SC-L).

The specific settings of the SC-Net architecture modules are shown in Table 1. The network contains four main stages, each performing feature extraction and processing at different resolution levels. In the 4× downsampling stage, the network first establishes basic feature representations through standard convolution operations. Starting from the 8× downsampling stage, The TFT modules are introduced for multi-type feature fusion, while the TFT continues to be applied in the 16× and 32× stages to process deeper semantic features.

Table 1. SC-net architecture module settings.

Stage	Stage 1	Stage 2	Stage 3	Stage 4
4×	1×1, 64 3×3, 64 1×1, 256	$1 \times b_1 \times m_1$ **TFT** h_1, w_{c1}, g_1	**TFT** h_2, w_{c2}, g_2	**TFT** h_3, w_{c3}, g_3
8×		3×3, N_{c1} 3×3, N_{c1}	$\times b_2 \times m_2$ 3×3, N_{c1} 3×3, N_{c1}	$\times b_4 \times m_4$ 3×3, N_{c1} 3×3, N_{c1}
16×			3×3, N_{c2} 3×3, N_{c2}	$\times b_3 \times m_3$ 3×3, N_{c2} 3×3, N_{c2}
32×				3×3, N_{c3} 3×3, N_{c3}

The parameters in Table 1 are defined as follows: (h_1, h_2, h_3) represents the number of heads in the TFT; (N_{c1}, N_{c2}, N_{c3}) represents the number of output channels; (b_1, b_2, b_3, b_4) represents the number of HRNet blocks; (m_1, m_2, m_3, m_4) represents the number of HRNet modules; (g_1, g_2, g_3) represents the number of feature slices in SA; and (w_{c1}, w_{c2}, w_{c3}) represents the width of cross-rectangular windows.

To accommodate different computational resources and accuracy requirements, the specific module parameter configurations of the three architecture variants are shown in Table 2.

Table 2. Module parameters for three architecture variants.

Variant	(b_1, b_2, b_3, b_4)	(m_1, m_2, m_3, m_4)	(g_1, g_2, g_3)	(w_{c1}, w_{c2}, w_{c3})	(h_1, h_2, h_3)	(c_1, c_2, c_3, c_4)
SC-T	(4,4,4,4)	(1,1,4,3)	(8,8,8)	(2,2,2)	(4,4,4)	(32,64,128,256)
SC-S	(4,4,4,4)	(1,1,4,3)	(9,9,9)	(4,4,4)	(4,4,4)	(36,72,144,288)
SC-L	(4,4,4,4)	(1,1,4,3)	(10,10,10)	(4,4,4)	(8,8,8)	(40,80,160,320)

SC-T, as the most lightweight variant, adopts smaller numbers of feature slices (8) and cross-window width (2), suitable for scenarios with limited computational resources. SC-S increases the number of feature slices (9) and window width (4), providing a balance between accuracy and efficiency. SC-L, as the largest variant, uses the most feature slices (10) and attention heads (8), achieving optimal segmentation performance but requiring more computational resources.

This hierarchical architecture design enables SC-Net to flexibly adapt to different application scenarios. Users can select appropriate variants based on specific performance requirements and computational constraints, finding the optimal balance between accuracy and efficiency.

4. Experiments

To validate the effectiveness of SC-Net in addressing the problems of insufficient spatial structure modeling and cross-scale information loss in remote sensing image semantic segmentation, we conducted extensive experiments on three classic remote sensing image datasets, including ablation studies, architecture variant comparisons, and comparative analysis with other state-of-the-art methods.

4.1. Datasets

The WHDLD dataset [53] is a high-density annotated remote sensing dataset created by Wuhan University, containing 4940 high-resolution remote sensing images of size 256×256 with RGB imaging bands, covering six categories: bare soil, buildings, pavement, vegetation, roads, and water bodies. The building and road categories in this dataset have obvious spatial geometric features, making it very suitable for validating our spatial structure modeling approach.

The ISPRS Potsdam dataset [54] comes from Potsdam, Germany, consisting of 38 images of size 6000×6000 with a resolution of 5cm, containing 6 categories. We selected partial images (Image IDs: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_14, 6_15, 7_13) and divided them into 256×256 size, totaling 6877 images. The high-resolution characteristics of this dataset make it an ideal choice for testing cross-scale information preservation capabilities.

The ISPRS Vaihingen dataset is a classic rural remote sensing dataset containing 3 areas and 33 images of different sizes with a resolution of 9 cm, bands including near-infrared, red, and green, and 6 land cover categories. We divided it into 12,491 images of size 256×256 for experimental research and removed background categories that had minimal impact on research significance.

To evaluate model robustness, we apply data augmentation techniques to all three datasets, including rotation, flipping, and Gaussian noise. Unless otherwise specified, all experiments use an 8:1:1 ratio for training, validation, and test sets.

4.2. Experimental setup and evaluation metrics

We employ the SGD optimizer during model training with a momentum of 0.9 and weight decay of 0.0001. The initial learning rate is set to 0.001 with a warmup strategy for the first 5 epochs, where the learning rate increases linearly from 0.0001 to 0.001. After warmup, we adopt a step decay strategy that halves the learning rate every 20 epochs, with the minimum learning rate set to 1e-6 and maximum epoch set to 100. The batch size is set to 16, and all training is performed on 2 GeForce RTX3090 GPUs with CUDA 11.3 and PyTorch 1.12.0. We utilize mixed precision training (FP16) with automatic mixed precision (AMP) to accelerate training and reduce memory consumption. The model employs dropout with a rate of 0.1 in the transformer modules to prevent overfitting. Data augmentation techniques include random rotation with angles (0°, 90°, 180°, 270°), horizontal and vertical flipping with a probability of 0.5, and Gaussian noise with a standard deviation of 0.01. To prevent statistical errors, all experiments are performed 3 times and the average of the 3 results is taken as the final result.

The three most representative evaluation metrics in semantic segmentation research are used in this paper: mean intersection over union (MIOU), mean pixel accuracy (MPA), and mean F1 score (MF1). These metrics can be obtained from the confusion matrix of the experimental results, which consists of four components: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The three evaluation indicators can be expressed as follows:

$$IOU = \frac{TP}{TP + FP + FN} \tag{4.1}$$

$$Precision = TP/(TP + FP) \tag{4.2}$$

$$Recall = TP/(TP + FN) \tag{4.3}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.4}$$

MIOU, MF1, and MPA represent the mean of all classes IOU, F1, and precision.

4.3. Architecture variant performance comparison

To validate the performance of SC-Net architecture variants, we conducted comparative experiments on the WHDLD dataset, comparing them with classic Deeplab models and HRNet with corresponding parameters. Table 3 shows the detailed experimental results.

From the results, SC-Net's FLOPs are significantly lower than those of HRNet with similar parameters, mainly due to the CSWin cross-window method adopted in the TFT that reduces computational complexity. Under these circumstances, SC-Net significantly outperforms HRNet with the same parameters, indicating that our designed feature attention layer and TFT can effectively integrate different types of features to achieve higher-quality feature encoding.

It is worth noting that DeepLabv3+ shows significant performance degradation when the parameter count decreases, possibly because atrous convolutions tend to lose more features in lightweight models. ABCNet enhances remote sensing image segmentation capability through dual-path focusing on spatial

and contextual information, but its performance is inferior to SC-Net, possibly due to a lack of fusion for different types of features to obtain high-level semantic features from multiple perspectives.

Table 3. Results of different architecture variants on the WHDLD dataset.

Method	Backbone	Params (M)	FLOPs (G)	MIOU	MPA	MF1
DeepLabv3+	ResNet34	22.4	7.932	54.07	67.08	67.51
DeepLabv3+	ResNet50	26.67	9.233	55.94	68.22	69.42
DeepLabv3+	ResNet101	45.67	14.108	58.64	71.20	71.94
ABCNet	ResNet34	23.72	6.579	58.01	69.01	70.10
ABCNet	ResNet50	28.79	9.693	60.99	72.15	73.01
ABCNet	ResNet101	47.78	14.712	62.17	74.21	75.82
HRNet	W32	29.55	11.339	56.87	70.01	70.32
HRNet	W36	37.27	13.982	58.43	71.11	71.56
HRNet	W40	45.89	16.862	61.09	72.89	73.68
SC-Net	SC-T	29.33	10.507	58.27	70.26	71.27
SC-Net	SC-S	37.14	12.327	61.33	73.07	73.89
SC-Net	SC-L	45.68	14.618	63.04	75.54	76.22

4.3.1. Overall architecture ablation experiment

To validate the effectiveness of the proposed modules, we conducted ablation experiments on the WHDLD dataset. We used the version with a TFT, FAL, and decoder modules removed as the baseline. Table 4 and Figure 7 show the evaluation metric results and partial visualization prediction results of the experiments, respectively.

Table 4. Overall architecture ablation study results on the WHDLD dataset.

Baseline	TFT	FAL	SC-up	HR-up	Params (M)	FLOPs (G)	MIOU	MPA	MF1
✓	✓			✓	45.09	13.231	62.10	73.67	74.40
✓	✓		✓		45.25	13.389	62.49	74.53	75.19
✓	✓	✓	✓		45.68	14.618	63.04	75.54	76.22
HRNet(W40)						16.862	61.09	72.89	73.68

The ablation results reveal several important insights about module interactions. First, adding a TFT alone (first row) improves MIOU by 0.39% over the baseline while reducing FLOPs, demonstrating its efficiency in feature fusion without computational overhead. Second, the decoder module (second row) contributes an additional 0.39% improvement, showing that progressive upsampling effectively preserves spatial details. Third, the complete model with FAL (third row) achieves the highest performance with 63.04% MIOU, indicating that FAL's spatial and multi-scale attention mechanisms provide complementary benefits to the TFT's feature fusion. The consistent performance gains at each stage suggest that the three modules address different aspects of the segmentation task without redundancy: the TFT handles heterogeneous feature fusion, the decoder manages spatial resolution recovery, and FAL enhances specific feature types. Notably, we observe no negative interactions among modules, as each addition monotonically improves performance. The visualization in Figure 7 further confirms

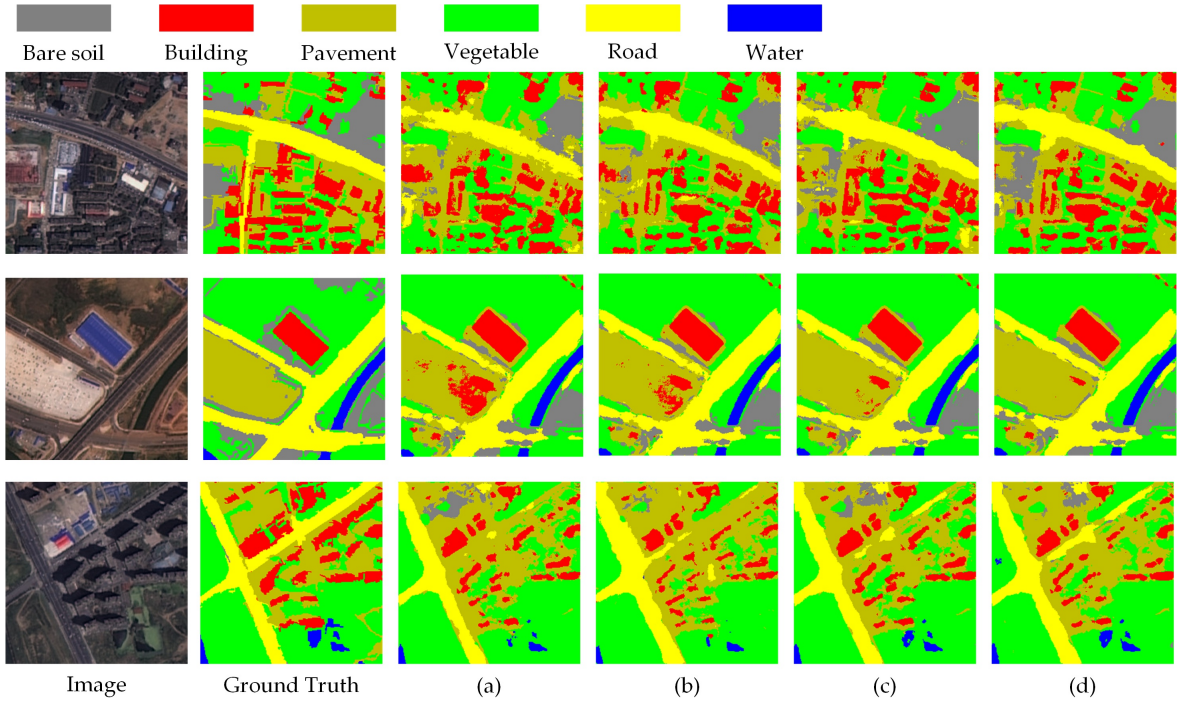

Figure 7. Ablation study visualization results showing progressive improvements with each module addition.

these quantitative findings, where each module addition progressively refines boundary accuracy and reduces misclassification, particularly in challenging regions marked by red boxes.

4.4. Ablation studies

Figure 7 shows the visualization results of the ablation experiments. We can see that after adding each module, the prediction results are closer to the ground truth labels, demonstrating the effectiveness of our designed modules. Particularly in building boundaries and small target areas, the advantages of SC-Net in spatial structure modeling and cross-scale information preservation are clearly visible.

4.4.1. Feature attention layer ablation study

In this section, we compare the attention modules used in our feature attention layer with other excellent module combinations, including CBAM+CA and DAN+ECA, where CBAM and DAN use two branches to obtain spatial features and channel features from different perspectives, CA groups features to model feature directional information, and ECA learns relationships between local and global features through feature compression.

The results in Table 5 show that our designed SA+MA combination achieves better performance. We believe the reason is that the four-branch different-scale pooling layers of MA can better combine with SC-Net's multi-branch structure to extract more detailed multi-scale features; SA hierarchically processes features, focusing on spatial feature information encoding modeling that might be ignored by other models, helping to further enhance network segmentation performance.

To gain deeper insights into the operational mechanisms of different modules, we employed Grad-

CAM visualization to analyze feature activation maps in building category prediction results (as shown in Figure 8). The visualization results clearly demonstrate that our proposed SA+MA architecture achieves superior performance compared to DAN+ECA and CBAM+CA configurations in building contour localization. Notably, in building boundary regions, SA+MA produces more concentrated and distinct activation responses, attributed to the effective spatial structure information modeling of the SA module. Additionally, through the multi-scale feature extraction capability of the MA module, the model exhibits enhanced adaptability to building targets of different scales.

Table 5. Feature attention layer ablation study results on the WHDLD and Potsdam datasets.

FAL	Params (M)	FLOPs (G)	WHDLD			Potsdam		
			MIOU	MPA	MF1	MIOU	MPA	MF1
DAN+ECA	45.74	14.735	61.83	73.88	74.26	70.54	81.94	81.02
CBAM+CA	45.64	14.790	62.07	74.01	74.31	70.23	82.01	81.45
SA+MA	45.68	14.618	63.04	75.54	76.22	71.57	83.02	82.77

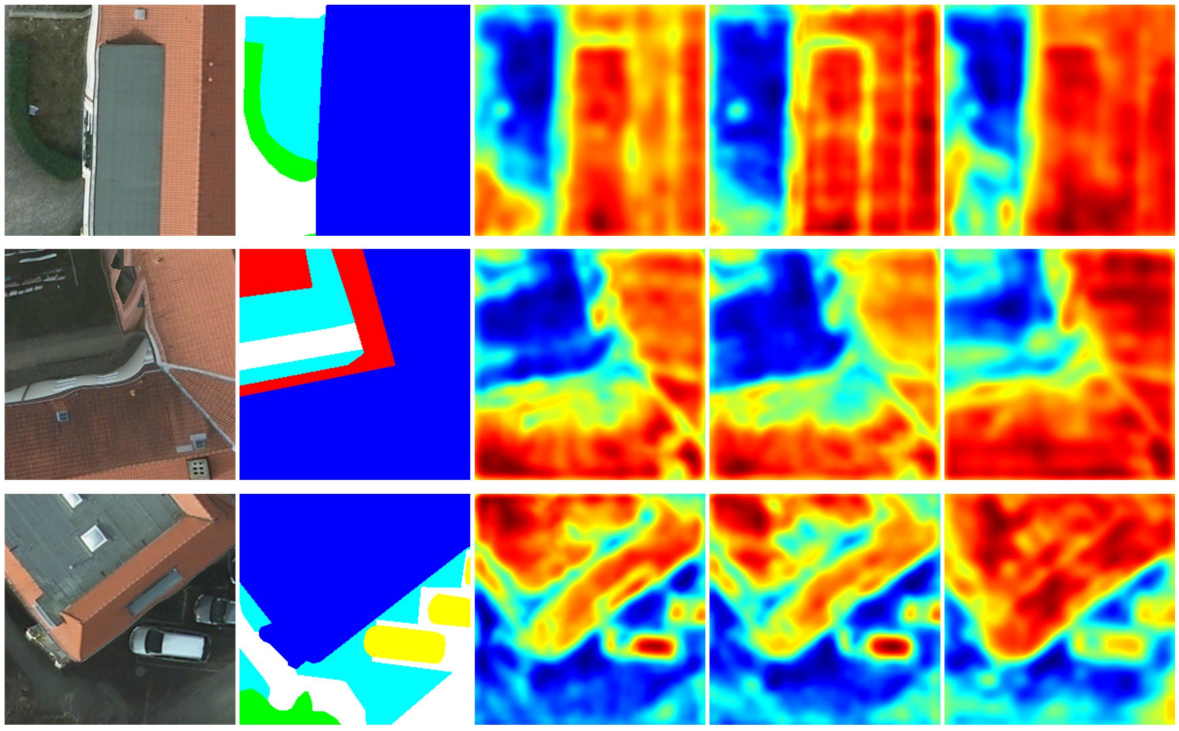

Figure 8. Grad-CAM visualization comparison of different attention mechanisms for building category prediction.

4.4.2. Window partitioning method ablation study

In this section, we conducted ablation experiments on our designed window partitioning. Table 6 shows partitioning in Swin-T square windows and CSwin-T cross-rectangular windows, respectively. We can see that the model performs better with the latter method, possibly because cross-rectangular

windows can focus on areas around feature points through different heads, having advantages in global modeling effects, while square windows may need to rely on more layers to increase global modeling capability. However, too many layers lead to high hardware consumption and model overfitting.

Table 6. Window partitioning method ablation study results on the WHDLD and Potsdam datasets.

Partitioning	$w_s \times w_s / w_c \times H$	WHDLD			Potsdam		
		MIOU	MPA	MF1	MIOU	MPA	MF1
Swin-T	16×16	62.17	74.31	74.88	70.77	82.32	81.34
CSWin-T	4×64	63.04	75.54	76.22	71.57	83.02	82.77

4.4.3. Hyperparameter sensitivity analysis

To validate our design choices, we conduct sensitivity analysis on key architectural parameters. Table 7 shows the impact of feature slice number g in SA and cross-window width w_c in the TFT on the WHDLD dataset using the SC-L variant.

For the feature slice number g, performance improves as g increases from 6 to 9, with MIOU rising from 61.87% to 63.04%. However, further increasing g to 10 yields only marginal gains of 63.09%, suggesting that 9 slices provide sufficient directional feature decomposition for capturing spatial geometric patterns. The cross-window width w_c demonstrates optimal performance at $w_c = 4$ with 63.04% MIOU. Smaller windows of $w_c = 2$ achieve only 62.17% MIOU, lacking sufficient contextual information for effective global modeling, while larger windows of $w_c = 8$ reach 62.83% MIOU but increase computational complexity without proportional performance benefits. These results confirm the robustness of our default parameter settings and provide practical guidance for applying SC-Net to different remote sensing scenarios.

Table 7. Sensitivity analysis of key hyperparameters on the WHDLD dataset.

Parameter setting	MIOU (%)	MPA (%)
Feature slices g in SA (with $w_c = 4$)		
$g = 6$	61.87	73.21
$g = 8$	62.76	74.88
$g = 9$ (default)	**63.04**	**75.54**
$g = 10$	63.09	75.61
Cross-window width w_c in TFT (with $g = 9$)		
$w_c = 2$	62.17	74.31
$w_c = 4$ (default)	**63.04**	**75.54**
$w_c = 8$	62.83	75.29

4.5. Comparison with other methods

4.5.1. Comparison on the WHDLD dataset

We conducted SC-Net experiments on the WHDLD dataset under different training set ratios and compared with other excellent works, including: AerialFormer [55], SERNet [56], HRViT [57],

DecoupleNet [58], SFA-Net [59], A2-FPN [60], UNetFormer, HRFormer [61], and RSSFormer.

Table 8 shows experimental results under different dataset split ratios (training:validation:test). The results demonstrate that SC-Net consistently outperforms other methods under different dataset splits. Specifically, under the 8:1:1 split ratio, SC-Net achieved 63.04% MIOU, surpassing transformer-based state-of-the-art methods such as AerialFormer and SERNet.

Table 8. Comparison results under different dataset split ratios on the WHDLD dataset.

Method	Backbone	8:1:1			7:1:2			6:2:2		
		MIOU	MPA	MF1	MIOU	MPA	MF1	MIOU	MPA	MF1
HRViT	HRViT-b3	60.78	73.89	73.72	60.11	72.28	72.60	59.01	71.99	72.23
HRFormer	HRFormer-B	61.01	74.22	74.56	60.43	72.57	73.36	59.88	72.07	73.72
A2-FPN	ResNet101	61.22	74.43	74.99	60.77	74.01	74.09	60.17	73.71	73.32
SERNet	EResNet	61.47	73.56	74.32	60.35	72.47	73.88	59.56	71.01	72.48
DecoupleNet	DNetD2	61.54	73.95	73.95	60.83	72.22	73.57	60.31	71.45	70.34
UNetFormer	ResNet101	61.56	74.02	74.22	60.80	73.76	74.13	60.24	73.03	73.81
RSSFormer	RSS-L	61.94	74.46	74.87	61.17	75.15	74.99	60.69	73.51	74.02
SFA-Net	EfficientNet	62.32	73.97	73.48	61.45	72.46	72.94	59.87	71.65	71.46
AerialFormer	ViT-B	62.37	72.99	73.54	62.01	74.36	72.57	60.01	71.47	72.35
SC-Net	SC-L	**63.04**	**75.54**	**76.22**	**62.37**	**75.24**	**75.42**	**61.45**	**74.00**	**74.83**

This improvement can be attributed to our efficient multi-type feature fusion strategy. While AerialFormer excels at capturing global dependencies through its aerial perspective design, it lacks effective mechanisms for integrating spatial structural information. HRViT and HRFormer, both built on high-resolution backbone networks, achieve competitive performance but show limitations in feature fusion.

The visualization results in Figure 9 further confirm our quantitative findings, showing that SC-Net produces more precise segmentation boundaries and better preserves details compared to other methods. This is particularly evident in challenging areas with significant scale variations or complex spatial structures.

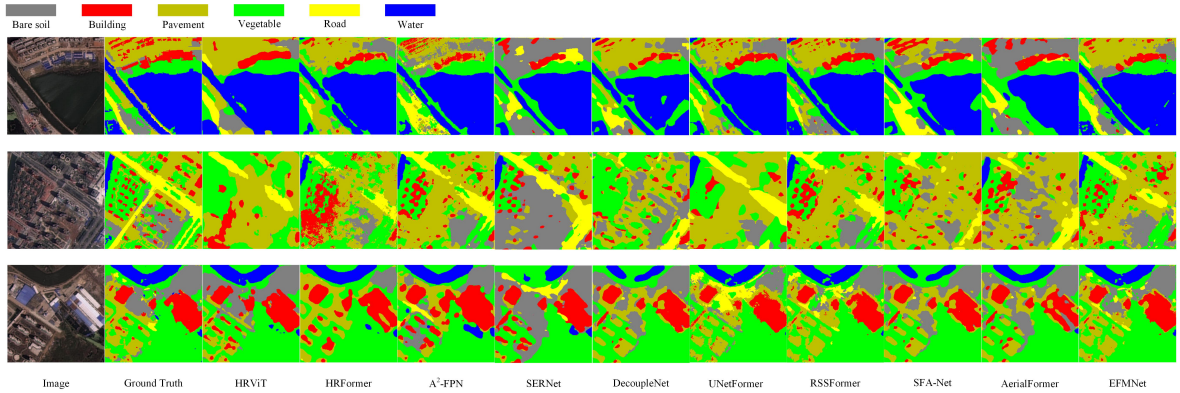

Figure 9. Visualization comparison results on the WHDLD dataset showing SC-Net's superior boundary preservation and detail recovery.

4.5.2. Comparison on Potsdam and Vaihingen datasets

To validate SC-Net's generalization capability across different scenarios, we conducted experiments on the Potsdam and Vaihingen datasets, which represent different remote sensing image scenarios. We compared with other excellent works, including AerialFormer, SERNet, DecoupleNetD2, SFA-Net, DAFormer [62], ST-UNet [63], SegViT [64], DC-Swin [65], and ABCNet.

Table 9 shows the results of these experiments. We can see that better results were obtained on both datasets, proving that SC-Net can handle datasets from different scenarios well. Compared to CNN-based models such as SERNet and SFA-Net, SC-Net can more effectively capture global feature information. Compared to DC-Swin, SC-Net's multi-branch design allows it to extract diverse feature types, enabling the model to obtain advanced semantic representations.

Table 9. Comparison results on the Vaihingen and Potsdam datasets.

Method	Backbone	Vaihingen			Potsdam		
		MIOU	MPA	MF1	MIOU	MPA	MF1
DAFormer	–	79.23	87.36	87.81	67.79	80.01	80.12
ST-UNet	–	79.42	87.31	87.73	68.33	79.23	79.45
SegViT	ViT-B	79.01	87.22	88.24	69.01	79.57	80.01
DecoupleNet	DNetD2	79.45	86.32	86.45	69.32	81.25	80.02
UNetFormer	ResNet101	79.29	87.34	88.42	69.45	81.32	81.21
SERNet	EResNet	79.49	87.01	87.64	70.01	81.49	81.35
DC-Swin	Swin-B	79.67	87.37	88.01	69.67	80.19	80.74
SFA-Net	EfficientNet	80.37	86.23	87.45	70.01	81.37	80.49
AerialFormer	ViT-B	80.45	85.94	87.21	69.90	82.01	81.22
SC-Net	SC-L	**81.57**	**88.82**	**89.42**	**71.57**	**83.02**	**82.77**

Figure 10 shows the visualization comparison results on the Potsdam dataset. We can see that SC-Net's results are close to the ground truth labels, performing better in the "building" category. We believe this may be due to our designed SA having stronger attention to spatial categories.

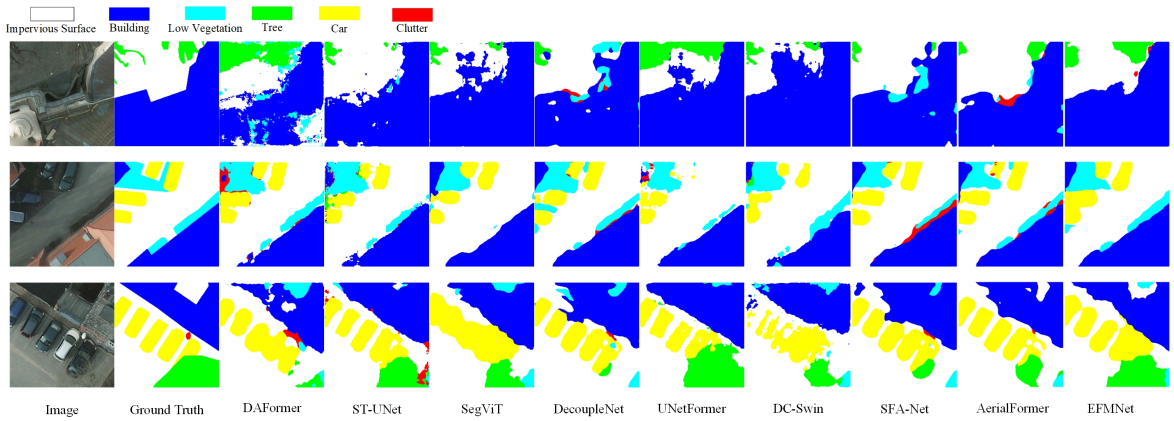

Figure 10. Visualization comparison results on Potsdam dataset demonstrating SC-Net's effectiveness in building segmentation.

4.6. Computational efficiency analysis

To provide a comprehensive evaluation of model efficiency, we report the training and inference time for all SC-Net variants on the WHDLD dataset. Table 10 summarizes the computational performance metrics.

Table 10. Computational efficiency comparison of SC-Net variants on WHDLD dataset.

Model	Params (M)	FLOPs (G)	Training time (s/epoch)	Inference time (ms/image)	FPS
SC-T	29.33	10.507	165	18.4	54.3
SC-S	37.14	12.327	192	22.7	44.1
SC-L	45.68	14.618	223	27.5	36.4
HRNet-W32	29.55	11.339	178	21.3	46.9
HRNet-W40	45.89	16.862	248	32.8	30.5
DeepLabv3+	26.67	9.233	152	16.2	61.7

As shown in Table 10, SC-Net achieves a favorable balance between accuracy and efficiency. SC-T demonstrates competitive training efficiency at 165 seconds per epoch (approximately 2.75 minutes) and achieves 54.3 FPS during inference with 18.4 ms per image, which is suitable for real-time applications. Although slightly slower than DeepLabv3+ (152 s/epoch), SC-T achieves significantly higher accuracy (58.27% vs. 55.94% MIOU from Table 3). Notably, SC-L demonstrates better computational efficiency than HRNet-W40 with comparable parameters, requiring 10.1% less training time per epoch and 16.2% faster inference speed, while achieving 1.95% higher MIOU. This efficiency gain is primarily attributed to the cross-rectangular window mechanism in the TFT that reduces computational complexity compared to full attention mechanisms.

5. Conclusions

5.1. Main contributions

Our core contributions include: First, we designed the FAL module, where SA specifically enhances modeling capabilities for spatial feature categories such as buildings through feature slicing, and MA reduces small-scale feature loss through multi-scale pooling and learnable feature storage units. Second, we proposed the TFT architecture, which achieves efficient fusion of spatial, multi-scale, and global features based on cross-rectangular window partitioning and nine-group key-value pair interaction mechanisms, reducing computational complexity while establishing explicit correlation relationships between heterogeneous features. Additionally, the multi-branch cascaded decoder adopts a progressive feature reconstruction strategy, effectively improving detail recovery capabilities.

5.2. Performance and limitations

Our proposed SC-Net achieves state-of-the-art performance across three benchmark datasets. Specifically, SC-Net-L attains 63.04% MIOU, 75.54% MPA, and 76.22% MF1 on the WHDLD dataset, representing improvements of 0.67-2.12% MIOU over existing methods. On the ISPRS Potsdam dataset, we achieve 71.57% MIOU, 83.02% MPA, and 82.77% MF1. For the ISPRS Vaihingen dataset, our

method reaches 81.57% MIOU, 88.82% MPA, and 89.42% MF1, demonstrating consistent superiority across different remote sensing scenarios.

However, several limitations should be acknowledged. First, while SC-Net shows strong performance on regular geometric objects like buildings and roads, it may be less effective for highly irregular natural objects such as complex vegetation boundaries or irregular water bodies. Second, the model's computational cost, though optimized through cross-rectangular windows, remains higher than lightweight CNN-based methods, which may limit deployment on resource-constrained platforms. Third, our experiments focus on images with 256×256 resolution, and the model's scalability to very high-resolution images (e.g., 2048×2048) requires further investigation. Fourth, the current design relies on three separate modules (FAL, TFT, decoder), and exploring more integrated architectures could potentially reduce parameter redundancy. Finally, while we validate on three datasets, testing on more diverse remote sensing scenarios (e.g., multi-temporal analysis, different sensor types) would further establish the method's generalizability.

Future research will focus on addressing these limitations by developing more lightweight variants for edge deployment, extending the framework to handle irregular objects more effectively, and exploring its application in multi-temporal change detection and cross-sensor scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. L. Wang, B. Zuo, Y. Le, Y. Chen, J. Li, Penetrating remote sensing: next-generation remote sensing for transparent earth, *Innovation*, **4** (2023), 100519. https://doi.org/10.1016/j.xinn.2023.100519
2. L. Huang, B. Jiang, S. Lv, Y. Liu, Y. Fu, Deep learning-based semantic segmentation of remote sensing images: a survey, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **17** (2024), 8370–8396. https://doi.org/10.1109/JSTARS.2023.3335891
3. Z. Che, L. Shen, L. Huo, C. Hu, Y. Wang, Y. Lu, et al., MAFF-HRNet: multi-attention feature fusion HRNet for building segmentation in remote sensing images, *Remote Sens.*, **15** (2023), 1382. https://doi.org/10.3390/rs15051382
4. M. Li, Y. Chen, T. Zhang, W. Huang, TA-YOLO: a lightweight small object detection model based on multi-dimensional trans-attention module for remote sensing images, *Complex Intell. Syst.*, **10** (2024), 5459–5473. https://doi.org/10.1007/s40747-024-01448-6
5. X. Yu, S. Li, Y. Zhang, Incorporating convolutional and transformer architectures to enhance semantic segmentation of fine-resolution urban images, *Eur. J. Remote Sens.* , **57** (2024), 2361768. https://doi.org/10.1080/22797254.2024.2361768
6. W. Hua, Q. Chen, A survey of small object detection based on deep learning in aerial images, *Artif. Intell. Rev.*, **58** (2025), 162. https://doi.org/10.1007/s10462-025-11150-9

7. F. Wang, Y. Zhang, Q. Hu, Y. Zhu, Remote sensing image semantic segmentation network based on multi-scale feature enhancement fusion, *Geocarto Int.*, **39** (2024), 2297330. https://doi.org/10.1080/10106049.2024.2375585
8. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
9. O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, Cham, **9351** (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
10. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science()*, Springer, Cham, **11211** (2018), 801–818. https://doi.org/10.1007/978-3-030-01234-2_49
11. Z. Luo, J. Pan, Y. Hu, L. Deng, Y. Li, C. Qi, et al., RS-Dseg: semantic segmentation of high-resolution remote sensing images based on a diffusion model component with unsupervised pretraining, *Sci. Rep.*, **14** (2024), 18609. https://doi.org/10.1038/s41598-024-69022-1
12. Z. Li, T. Qu, Q. Chong, J. Xu, FMCNet: a fuzzy multiscale convolution network for remote sensing image segmentation, *Can. J. Remote Sens.*, **50** (2024), 2418091. https://doi.org/10.1080/07038992.2024.2418091
13. W. Boulila, H. Ghandorh, S. Masood, A. Alzahem, A. Koubaa, F. Ahmed, et al., A transformer-based approach empowered by a self-attention technique for semantic segmentation in remote sensing, *Heliyon*, **10** (2024), e29396. https://doi.org/10.1016/j.heliyon.2024.e29396
14. X. Wang, H. Wang, Y. Jing, X. Yang, J. Chu, A bio-inspired visual perception transformer for cross-domain semantic segmentation of high-resolution remote sensing images, *Remote Sens.*, **16** (2024), 1514. https://doi.org/10.3390/rs16091514
15. S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 6881–6890.
16. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 12077–12090.
17. Y. L. Chen, C. L. Lin, Y. C. Lin, T. C. Chen, Transformer-CNN for small image object detection, *Signal Process. Image Commun.*, **129** (2024), 117194. https://doi.org/10.1016/j.image.2024.117194
18. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 10012–10022.
19. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, et al., CSwin transformer: a general vision transformer backbone with cross-shaped windows, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 12124–12134.

20. J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, et al., Deep high-resolution representation learning for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2020), 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686

21. S. Cheng, B. Li, L. Sun, Y. Chen, HRRNet: hierarchical refinement residual network for semantic segmentation of remote sensing images, *Remote Sens.*, **15** (2023), 1244. https://doi.org/10.3390/rs15051244

22. H. Wu, C. Liang, M. Liu, Z. Wen, Optimized HRNet for image semantic segmentation, *Expert Syst. Appl.*, **174** (2021), 114532. https://doi.org/10.1016/j.eswa.2020.114532

23. X. Yang, X. Fan, M. Peng, Q. Guan, L. Tang, Semantic segmentation for remote sensing images based on an AD-HRNet model, *Int. J. Digit. Earth*, **15** (2022), 2376–2399. https://doi.org/10.1080/17538947.2022.2159080

24. H. Feng, T. Zhong, Backbone feature enhancement and decoder improvement in HRNet for semantic segmentation, *Int. J. Adv. Comput. Sci. Appl.*, **15** (2024). https://doi.org/10.14569/ijacsa.2024.0151098

25. J. Xiang, J. Liu, D. Chen, Q. Xiong, C. Deng, CTFuseNet: a multi-scale CNN-transformer feature fused network for crop type segmentation on UAV remote sensing imagery, *Remote Sens.*, **15** (2023), 1151. https://doi.org/10.3390/rs15041151

26. J. Yang, H. Wan, Z. Shang, Enhanced hybrid CNN and transformer network for remote sensing image change detection, *Sci. Rep.*, **15** (2025), 10161.

27. Z. Zhang, L. Huang, B. H. Tang, W. Le, M. Wang, J. Cheng, et al., MATNet: multiattention transformer network for cropland semantic segmentation in remote sensing images, *Int. J. Digit. Earth*, **17** (2024), 2392845. https://doi.org/10.1080/17538947.2024.2392845

28. M. Liu, P. Liu, L. Zhao, Y. Ma, L. Chen, M. Xu, Fast semantic segmentation for remote sensing images with an improved short-term dense-connection (STDC) network, *Int. J. Digit. Earth*, **17** (2024), 2356122. https://doi.org/10.1080/17538947.2024.2356122

29. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19.

30. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, et al., Dual attention network for scene segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 3146–3154.

31. J. Lv, Q. Shen, M. Lv, Y. Li, L. Shi, P. Zhang, Deep learning-based semantic segmentation of remote sensing images: a review, *Front. Ecol. Evol.*, **11** (2023), 1201125. https://doi.org/10.3389/fevo.2023.1201125

32. X. Wang, Z. Hu, S. Shi, M. Hou, L. Xu, X. Zhang, A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved UNet, *Sci. Rep.*, **13** (2023), 7600. https://doi.org/10.1038/s41598-023-34379-2

33. X. Li, J. Li, MFCA-Net: a deep learning method for semantic segmentation of remote sensing images, *Sci. Rep.*, **14** (2024), 5745. https://doi.org/10.1038/s41598-024-56211-1

34. A. Yu, Y. Quan, R. Yu, W. Guo, X. Wang, D. Hong, et al., Deep learning methods for semantic segmentation in remote sensing with small data: a survey, *Remote Sens.*, **15** (2023), 4987. https://doi.org/10.3390/rs15204987

35. Y. Mo, Y. Wu, X. Yang, F. Liu, Y. Liao, Review the state-of-the-art technologies of semantic segmentation based on deep learning, *Neurocomputing*, **493** (2022), 626–646. https://doi.org/10.1016/j.neucom.2022.01.005

36. H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, C. Tao, SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images, *IEEE Geosci. Remote Sens. Lett.*, **18** (2020), 905–909. https://doi.org/10.1109/LGRS.2020.2988294

37. X. Li, F. Xu, F. Liu, X. Lyu, Y. Tong, Z. Xu, et al., A synergistical attention model for semantic segmentation of remote sensing images, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–16. https://doi.org/10.1109/TGRS.2023.3243954

38. R. Wang, L. Ma, G. He, B. A. Johnson, Z. Yan, M. Chang, et al., Transformers for remote sensing: a systematic review and analysis, *Sensors*, **24** (2024), 3495. https://doi.org/10.3390/s24113495

39. S. Paheding, A. Saleem, M. F. H. Siddiqui, N. Rawashdeh, A. Essa, A. A. Reyes, Advancing horizons in remote sensing: a comprehensive survey of deep learning models and applications in image classification and beyond, *Neural Comput. Appl.*, **36** (2024), 16727–16767. https://doi.org/10.1007/s00521-024-10165-7

40. L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, et al., UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, *ISPRS J. Photogramm. Remote Sens.*, **190** (2022), 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008

41. G. Liang, F. Xie, Y. R. Chien, Class-aware self-and cross-attention network for few-shot semantic segmentation of remote sensing images, *Mathematics*, **12** (2024), 2761. https://doi.org/10.3390/math12172761

42. F. Xie, G. Liang, Y. R. Chien, Global–local query-support cross-attention for few-shot semantic segmentation, *Mathematics*, **12** (2024), 2936. https://doi.org/10.3390/math12182936

43. W. Lu, Y. Hu, W. Shao, H. Wang, Z. Zhang, M. Wang, A multiscale feature fusion enhanced CNN with the multiscale channel attention mechanism for efficient landslide detection (MS2LandsNet) using medium-resolution remote sensing data, *Int. J. Digit. Earth*, **17** (2024), 2300731. https://doi.org/10.1080/17538947.2023.2300731

44. D. Lu, S. Cheng, L. Wang, S. Song, Multi-scale feature progressive fusion network for remote sensing image change detection, *Sci. Rep.*, **12** (2022), 11968. https://doi.org/10.1038/s41598-022-16329-6

45. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2117–2125.

46. J. Liu, H. Gu, Z. Li, H. Chen, H. Chen, Multi-scale feature fusion attention network for building extraction in remote sensing images, *Electronics*, **13** (2024), 923. https://doi.org/10.3390/electronics13050923

47. H. Zheng, M. Zhang, M. Gong, A. K. Qin, T. Liu, F. Jiang, Multi-scale hierarchical feature fusion network for change detection, *Pattern Recognit.*, **161** (2025), 111266. https://doi.org/10.1016/j.patcog.2024.111266

48. C. Wang, L. Li, Z. Wang, J. Ma, Y. Kong, Y. Wang, et al., Multi-scale dense graph attention network for hyperspectral classification, *Can. J. Remote Sens.*, **50** (2024), 2333424. https://doi.org/10.1080/07038992.2024.2333424

49. Y. Liu, K. Gao, H. Wang, Z. Yang, P. Wang, S. Ji, et al., A transformer-based multi-modal fusion network for semantic segmentation of high-resolution remote sensing imagery, *Int. J. Appl. Earth Obs. Geoinf.*, **133** (2024), 104083. https://doi.org/10.1016/j.jag.2024.104083

50. H. Wang, H. Wang, L. Wu, TGF-Net: transformer and gist CNN fusion network for multi-modal remote sensing image classification, *PLoS One*, **20** (2025), e0316900. https://doi.org/10.1371/journal.pone.0316900

51. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155

52. Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 13708–13717. https://doi.org/10.1109/CVPR46437.2021.01350

53. Z. Shao, W. Zhou, X. Deng, M. Zhang, Q. Cheng, Multilabel remote sensing image retrieval based on fully convolutional network, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **13** (2020), 318–328. https://doi.org/10.1109/JSTARS.2019.2961634

54. F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, et al., The ISPRS test project on urban classification and 3D building reconstruction, *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, **I-3** (2012), 293–298. https://doi.org/10.5194/isprsannals-I-3-293-2012

55. T. Hanyu, K. Yamazaki, M. Tran, R. A. McCann, H. Liao, C. Rainwater, et al., AerialFormer: multi-resolution transformer for aerial image segmentation, *Remote Sens.*, **16** (2024), 2930. https://doi.org/10.3390/rs16162930

56. X. Zhang, L. Li, D. Di, J. Wang, G. Chen, W. Jing, et al., SERNet: squeeze and excitation residual network for semantic segmentation of high-resolution remote sensing images, *Remote Sens.*, **14** (2022), 4770. https://doi.org/10.3390/rs14194770

57. J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y. H. Chen, et al., Multi-scale high-resolution vision transformer for semantic segmentation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 12084–12093. https://doi.org/10.1109/CVPR52688.2022.01178

58. Y. Zhang, J. Yang, Y. Liu, J. Tian, S. Wang, C. Zhong, et al., Decoupled pyramid correlation network for liver tumor segmentation from CT images, *Med. Phys.*, **49** (2022), 7207–7221. https://doi.org/10.1002/mp.15723

59. X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, et al., Semantic flow for fast and accurate scene parsing, in *Computer Vision – ECCV 2020. Lecture Notes in Computer Science()*, Springer, Cham, **12346** (2020). https://doi.org/10.1007/978-3-030-58452-8_45

60. R. Li, L. Wang, C. Zhang, C. Duan, S. Zheng, A2-FPN for semantic segmentation of fine-resolution remotely sensed images, *Int. J. Remote Sens.*, **43** (2022), 1131–1155. https://doi.org/10.1080/01431161.2022.2030071

61. Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, et al., HRFormer: high-resolution transformer for dense prediction, preprint, arXiv:2110.09408. https://doi.org/10.48550/arXiv.2110.09408

62. L. Hoyer, D. Dai, L. V. Gool, DAFormer: improving network architectures and training strategies for domain-adaptive semantic segmentation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 9914–9925. https://doi.org/10.1109/CVPR52688.2022.00969

63. X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, Y. Xue, Swin transformer embedding UNet for remote sensing image semantic segmentation, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–15. https://doi.org/10.1109/TGRS.2022.3144165

64. B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, et al., SegViT: semantic segmentation with plain vision transformers, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 4971–4982.

65. L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, S. Fang, A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images, *IEEE Geosci. Remote Sens. Lett.*, **19** (2022), 1–5. https://doi.org/10.1109/LGRS.2022.3143368

AIMS Press