Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multiscale Context-Aware Network for Remote Sensing Images Semantic Segmentation

Ze Wang

xinghekmno4@gmail.com

Guizhou University

Jin Qin

Guizhou University

Chuhua Huang

Guizhou University

Yongjun Zhang

Guizhou University

Research Article

Keywords: Remote sensing images, Semantic segmentation, Convolutional neural network(CNN), Transformer

Posted Date: April 1st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6236533/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multiscale Context-Aware Network for Remote Sensing Image Segmentation

Ze Wang[1], Jin Qin[1], Chuhua Huang[1], Yongjun Zhang[1*]

[1*]The State Key Laboratory of Public Big Data, and the College of Computer Science and Technology, Guizhou University, Huaxi Avenue, Gui Yang, 550000, Gui Zhou, China.

*Corresponding author(s). E-mail(s): zyj6667@126.com;
Contributing authors: xinghekmno4@gmail.com; jqin1@gzu.edu.cn; chhuang@gzu.edu.cn;

Abstract

Remote sensing images typically exhibit significant scale and appearance variations, requiring deep learning-based semantic segmentation methods to effectively capture local and global contextual information to improve segmentation accuracy and handle complex scenes. Although CNN-based methods have been widely applied, they excel at capturing local details but are limited in modeling global context. By relying on multi-head self-attention, transformers can effectively capture global dependencies but often incur high computational costs. In this paper, we propose a Multiscale Context-Aware Network (MCANet) that combines CNN and Transformer architectures to comprehensively and efficiently model multi-scale local and global contextual relationships. Specifically, an Adaptive Feature Enhancement Module (AFEM) is designed to enhance local contextual representations of multiscale features in the encoder using large-kernel strip convolutions and frequency-adaptive weighting. Meanwhile, we develop a Multi-scale Global-context Transformer Block (MGTB) in the decoder to efficiently extract global contextual information across different scales. Furthermore, the Feature Fusion Module (FFM) is introduced to integrate the local context enhanced by AFEM and the global context generated by MGTB, thus further promoting the joint learning of local and global information. Extensive quantitative and qualitative experiments conducted on the public Vaihingen and Potsdam datasets demonstrate that the proposed MCANet achieves superior segmentation performance compared to existing mainstream methods.

Keywords: Remote sensing images, Semantic segmentation, Convolutional neural network(CNN), Transformer

1 Introduction

As an essential means of acquiring geospatial information, remote sensing technology plays a crucial role in urban planning, vegetation monitoring, military surveillance, disaster assessment, and meteorological forecasting. It has become an indispensable technical support in these areas. In recent years, with the rapid development of remote sensing technology, a large number of high-quality remote sensing images have been obtained, which contain rich spatial and semantic information. Semantic segmentation of remote sensing images enables the accurate assignment of class labels to each pixel, which is significant for deepening our understanding of the geographical environment.

Deep convolutional neural networks (DCNNs) have recently made significant progress in image segmentation tasks, mainly due to their ability to extract hierarchical features automatically. Fully Convolutional Network (FCN)[1], as the first deep learning network to achieve end-to-end pixel-level semantic segmentation, overcame the limitations of traditional methods. Subsequently, Ronneberger et al.[2] introduced an encoder-decoder structure and fused encoder features during the decoding phase using skip connections, better preserving spatial information. DeepLabV3+[3] introduced dilated convolutions and the Atrous Spatial Pyramid Pooling (ASPP) module, attempting to aggregate contextual information from different receptive fields and further enhancing the model's multiscale perception ability. Recent research has focused on capturing richer semantic information by incorporating attention mechanisms and combining multiscale features, enabling more refined segmentation of targets at different scales. For example, Cao et al.[4] proposed a Context Aggregation Network, which combines attention mechanisms with multiscale features, significantly improving target localization accuracy. Long et al.[5] proposed a threshold attention-based segmentation network, which effectively models the correlations between different regions of the feature map using a threshold attention mechanism and introduces threshold attention pyramid pooling to

facilitate the interaction between features at various scales.

Although these CNN-based methods have made significant progress in image semantic segmentation tasks, their reliance on local convolution operations results in small receptive fields, making it challenging to capture global contextual information fully. This limits their ability to model long-range dependencies in complex scenes. With its powerful long-range dependency modeling capability, the Vision Transformer (ViT)[6] has opened up new prospects for image processing. Unlike traditional CNN, ViT does not rely on local convolution operations but instead uses a global self-attention mechanism to capture dependencies between distant pixels in an image, thereby more effectively modeling long-range contextual information. Subsequently, Zheng et al.[7] proposed the SETR based on ViT, further advancing the application of transformers in semantic segmentation tasks. However, the multi-head self-attention (MSA) mechanism used by transformers suffers from the issue that its computational complexity grows quadratically with the scale of the input data, leading to higher computational overhead. To address this issue, the Swin Transformer[8] successfully models long-range dependencies while maintaining efficiency by restricting MSA to local windows and introducing a window-shifting mechanism. PVT[9] reduces the number of key-value pairs required in the traditional multi-head self-attention mechanism and uses a pyramid structure to extract image features from multiple scales, effectively reducing computational complexity. In addition, ResT[10] and CPVT[11] embed the relative and absolute position information of image patches into tokens through convolution operations, enabling the network to encode the positions of input images with different sizes adaptively, thus solving the limitation of ViT's ability to process only fixed-length sequences.

While transformers excel at capturing global context and modeling long-range dependencies, they often overlook the fine-grained features of local context, making them less effective than convolutional networks in handling local details. To tackle this issue, several studies have attempted to combine convolutional networks with transformers to capture both local and global context information. For instance, TransUNet[12] combines convolutional networks with the self-attention mechanism of transformers, enhancing long-range dependency modeling while preserving the advantages of convolutional networks in local feature extraction. ST-UNet[13] employs a parallel dual-encoder structure, integrating convolutional networks and Swin Transformer[8] to improve feature representation. UnetFormer[14] and CMLFormer[15] combine convolutional encoders with transformer decoders to achieve more effective local feature extraction and long-range dependency modeling. Zhang et al.[16] designed a remote sensing image segmentation network that combines Swin Transformer with convolutional decoder to improve model accuracy and robustness.

Inspired by the literature above, this paper proposes a novel hybrid architecture network, MCANet, which combines CNN and Transformer components to simultaneously capture local and global contextual information for accurate remote sensing image segmentation. Specifically, we use ResNet18 as the encoder to extract local contextual features efficiently. Given the limited receptive field of convolutional operations, we design an Adaptive Feature Enhancement Module (AFEM) that enhances feature representations at different scales by combining large convolution kernels and frequency-domain feature cues in parallel. To further improve the decoder's performance, we introduce a novel multiscale global context Transformer block, termed MGTB, which is designed to model long-range dependencies of features and capture multiscale global context. MGTB effectively captures global context through a window-based global context self-attention mechanism and replaces the traditional Transformer feed-forward network with depth-wise separable large convolutions at different scales, thus overcoming the long-range modeling issue caused by window limitations and optimizing the extraction of multi-scale contextual information. Additionally, we propose a Feature Fusion Module (FFM) to merge the local features from the encoder and the global features from the decoder, helping the model perceive both local and global contextual information at different scales. The main contributions of this paper are summarized as follows:

1. An AFEM module is designed to enhance the expressive power of features at each stage of the encoder.
2. We propose a novel MGTB module that effectively captures multi-scale global context in remote sensing images by combining window-based global context self-attention with a multi-scale context multi-layer perceptron.
3. An FFM module is introduced to fuse the local features from the encoder with the global features from the decoder, enhancing the joint representation of both local and global context information.
4. We propose a new MCANet network architecture, utilizing a lightweight CNN encoder and a multi-scale global context transformer decoder to effectively leverage both global and local context information for accurate segmentation of remote sensing images.

2 Related Work

2.1 Semantic segmentation of combining CNN and Transformer

CNN and Transformer each have unique advantages in image processing tasks. CNN excels at capturing local features due to its translation invariance and feature-sharing properties, making it highly effective for object detection at different scales. In contrast, Transformer is particularly good at modeling global dependencies and can reveal long-range semantic relationships in images. Recent studies[13–15] have shown that combining CNN with Transformer can efficiently fuse local and global information in image segmentation, improving performance. For example, DBCT-Net[17] adopts a

dual-branch hybrid CNN-Transformer network, combining CNN's local feature extraction with Transformer's global relationship modeling, achieving excellent performance. STransFuse[18] proposed a dual-encoder framework that utilizes CNN and Transformer to extract features and enhances feature representation through adaptive fusion. The CMT model[19] combines hierarchical Transformer and convolutional operations, balancing global and local feature extraction while reducing computational complexity. Based on these ideas, this paper proposes a novel architecture that integrates a CNN encoder with a Transformer-based global context decoder, effectively merging local and global information for more accurate remote sensing image segmentation.

2.2 Multiscale Context For Remote Sensing Images Segmentation

Objects in remote sensing images exhibit significant scale variations, which impose higher demands on the network's ability to capture multiscale contextual information. Accordingly, numerous approaches have been proposed to design multiscale feature extraction modules, typically by employing parallel convolutional layers with different receptive fields to model multiscale information. For instance, Wang et al.[20] proposed a Multi-feature Interactive Fusion Module (MIF-M), which establishes correlations among different features to achieve hierarchical feature fusion, thereby enhancing the network's perception and representation of multiscale contextual information. MACU-Net[21] leverages multiscale skip connections and asymmetric convolutional blocks to improve feature extraction and representation capabilities, thus acquiring rich contextual features. Chen et al.[22] combine a multiscale strip pooling attention module with a deep multi-head feedforward transfer attention module, enabling efficient extraction of multiscale contextual features. However, for large-scale objects that require broader receptive fields, directly employing such modules becomes inefficient because enlarging convolutional kernel sizes leads to exponential growth in model parameters and computational cost. Consequently, some studies attempt to enhance multiscale feature representation through cross-stage feature fusion. For example, CFNet[23] introduces a cross-fusion module to merge adjacent feature maps in the encoder, generating more diverse multiscale representations. MPSegNet[24] adopts a cascaded architecture that decomposes the overall segmentation task into three subnetworks, each focusing on small-scale, large-scale, and other-scale objects, thus progressively achieving accurate segmentation across scales and significantly improving the segmentation performance for objects of varying sizes. Unlike these methods, this work proposes AFEM and GCTB modules to capture both local and global contextual information effectively. Moreover, an FFM module is designed to integrate features from different stages of the encoder and decoder, enhancing the correlation among multiscale features and significantly improving the network's ability to represent and understand multiscale context, thereby advancing the overall segmentation performance in remote sensing imagery.

3 Methodology

3.1 Overall Architecture

The overall architecture of MCANet is shown in Fig.1, which adopts a classic encoder-decoder framework. The encoder is built upon the ResNet18[25] architecture to effectively extract low-level features from the input images. ResNet18 consists of four ResBlock stages, each performing 1/2 downsampling of the feature maps. The multiscale features extracted from the first three stages of the encoder are fed into the AFEM module for feature enhancement, and their outputs are connected to the decoder. The MGTB module is applied at different decoder stages to capture long-range dependencies. The enhanced features from AFEM and the output features from MGTB are fed into the FFM module to fuse multiscale contextual information effectively. Finally, the outputs of the first three stages of MGTB are upsampled and summed, then concatenated with the features from the final stage to fully leverage hierarchical feature representations and improve semantic segmentation accuracy. The concatenated features are passed through the segmentation head to generate the final prediction.

3.2 Adaptive Feature Enhancement Module

Due to the complexity of remote sensing images, small object segmentation is prone to confusion with the background. However, limited by the convolution operations, backbone networks typically have a narrow receptive field, making it difficult to distinguish small objects from the background. Given this, we propose the AFEM module, which enhances feature representation from two perspectives. First, from the perspective of expanding the receptive field, large-kernel strip convolutions are employed to capture broader spatial context and handle scale variations. Second, to differentiate between background and small object features, we enhance the global structural information by capturing low-frequency contextual features while strengthening edge and texture details through high-frequency information, thereby improving the model's discriminative capability for target regions. The structure of AFEM is illustrated in Fig.2. The input feature F_{in} is first normalized by a LayerNorm layer and then passed through a 1×1 convolution to adjust the channel dimension to twice its original size, thereby enhancing feature representation capability. The resulting high-dimensional feature $\tilde{F}$ is subsequently fed into two components: Spatial Context Extraction (SCE) and Adaptive Frequency Modulation (AFM).

In SCE, $\tilde{F}$ undergoes horizontal and vertical pooling to extract global information from different directions and generate corresponding feature vectors. These two feature vectors are then summed and broadcast to the original feature map resolution. Afterward, large-kernel

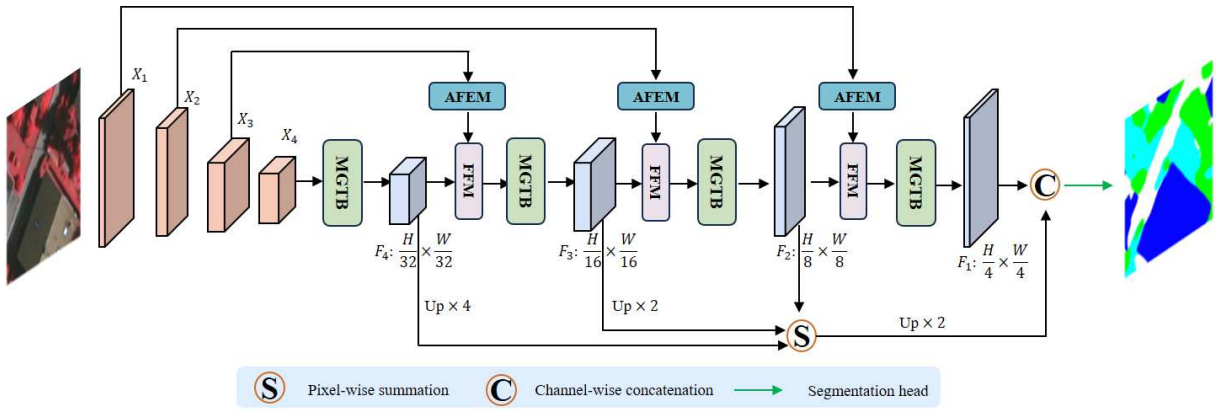


Fig. 1 Overall framework of the MCANet. The MCANet proposed in this paper uses ResNet18 as the backbone network for feature extraction, with the decoder consisting of the innovative MGTB and FFM modules. The AFEM module enhances the features from the first three stages of the encoder and then fuses with the output features from each stage of MGTB through the FFM module. Finally, the low-level features in the decoder undergo upsampling and summing operations, which are concatenated with the high-level features, and the final predicted output map is generated through the segmentation head.

strip convolutions with kernel sizes 1×7 and 7×1 are sequentially applied to capture multi-directional spatial context information. Finally, a 3×3 depthwise separable convolution is used to capture finer features.

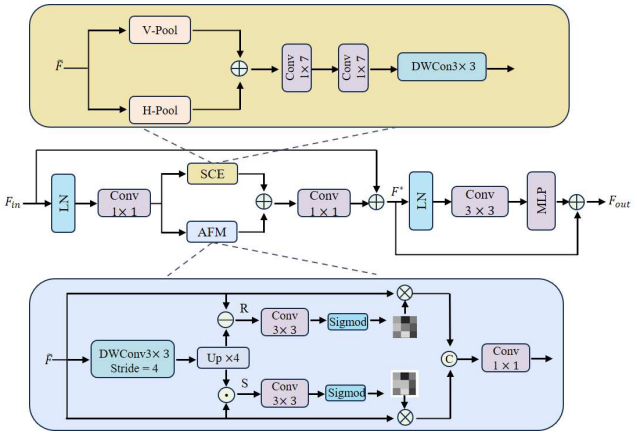


Fig. 2 Structure of the AFEM module. MLP consists of two layers of 1×1 convolutions.

In AFM, we adaptively adjust the weights of low-frequency background information and high-frequency detail features to model the target regions and suppress background interference finely. In image sharpening and contrast enhancement, high-frequency information is typically enhanced by calculating the difference between an image and its blurred version[26]. Inspired by this, we first downsample the input feature $\tilde{F}$ and obtain the blurred feature $\tilde{F}_b$ through linear interpolation. Then, the high-frequency detail features R are extracted by computing the difference between $\tilde{F}$ and $\tilde{F}_b$. For normalized images, low-frequency information can generally be obtained by pixel-wise multiplying the image with its blurred version. Following this idea, we perform element-wise multiplication between $\tilde{F}$ and $\tilde{F}_b$ to obtain the low-frequency feature S, which preserves global background information. Subsequently, both R and S are passed through separate 3×3 convolutions for further information extraction, and Sigmoid activation functions process the resulting features to generate corresponding weight maps, enabling adaptive modulation of high- and low-frequency information. Finally, the weighted high- and low-frequency features are concatenated along the channel dimension and passed through a 1×1 convolution to restore the channel dimension to match that of $\tilde{F}$, yielding the final frequency-modulated features.

The output features from SCE and AFM are summed, followed by a 1×1 convolution to adjust the dimensions to match the input feature F. Subsequently, a residual connection is introduced to enhance the feature representation, resulting in the feature F^*. Next, F^* is passed to the next stage, where it undergoes normalization, a series of convolution operations, and residual connections, ultimately generating the final enhanced feature representation.

3.3 Multiscale Global-Context Transformer Block

To accurately recognize complex categories, such as buildings at varying scales, winding rivers, and diverse vehicles in remote sensing images, it is crucial to efficiently leverage global context information. Standard Transformer modules typically rely on multi-head self-attention (MSA) and multi-layer perceptrons (MLP) to model global features. However, MSA has a computational complexity of $O(N^2)$, which leads to a significant increase in computational cost as the input image size grows. Additionally, MLP primarily operates along the channel dimension, which limits its ability to capture spatial relationships, thereby affecting the accurate recognition of complex categories. To address these challenges, we propose a low-complexity Multiscale Global-context Transformer Block (MGTB), as shown in Fig.3, designed to effectively extract global-context information from image regions at multiple scales. GCTB consists of two core components: Global-context Multi-Head Self-Attention (G-MSA) and Multiscale-context

Multi-Layer Perceptron (M-MLP). The G-MSA generates queries, keys, and values within local windows, computing attention matrices to capture local detail correlations in the image. Moreover, combining window-based horizontal and vertical pooling features to enable interactions between windows further strengthens the model's ability to capture spatial dependencies across different regions, enabling better incorporation of global context information. The M-MLP utilizes multiscale horizontal and vertical depthwise separable strip convolutions to enhance the model's ability to perceive features at different scales, improving its sensitivity to multiscale features and thereby optimizing semantic segmentation performance.

Specifically, for the input feature X, the MGTB first applies pixel attention (PA) for position encoding. Then, we perform a window partitioning operation to divide the features into n non-overlapping windows. To ensure that the feature map size is divisible by the specific window size, we set the window size to 8. Within each window, three linear layers are used to generate the query (Q), key (K), and value (V) vectors. For the i-th window, the input to the self-attention mechanism can be expressed as:

$$Q^i = X^i W_Q^i, \quad K^i = X^i W_K^i, \quad V^i = X^i W_V^i \tag{1}$$

The queries and keys of each window undergo a dot product to obtain the raw attention scores, which are then normalized using softmax to generate attention weights applied to the values. The computation is expressed as:

$$SA^i = \text{softmax}\left(\frac{Q^i (K^i)^T}{\sqrt{d^i}}\right) V, \quad i = 1, 2, \ldots, n \tag{2}$$

where SA^i represents the self-attention of the i-th window feature, and d^i is the channel dimension of Q^i and K^i.

In addition to calculating self-attention within each window, we introduce channel attention to allow MGTB to capture global information simultaneously from both spatial and channel dimensions. Specifically, each window undergoes a dot product operation to obtain the raw attention scores, which are then processed through average pooling and max pooling operations. The results of these operations are added together and passed through a sigmoid activation function to generate channel attention scores. The computation is expressed as:

$$CA^i = \sigma\left(P_{\text{avg}}(Q^i (K^i)^T) \oplus P_{\max}(Q^i (K^i)^T)\right) \tag{3}$$

where CA^i represents the channel attention for the i-th window feature, and $\oplus$ denotes element-wise summation. P_{avg} and $P_{\max}$ represent the average pooling and max pooling operations, respectively.

After obtaining the self-attention and channel attention for each window, the attention head of the i-th window can be expressed as:

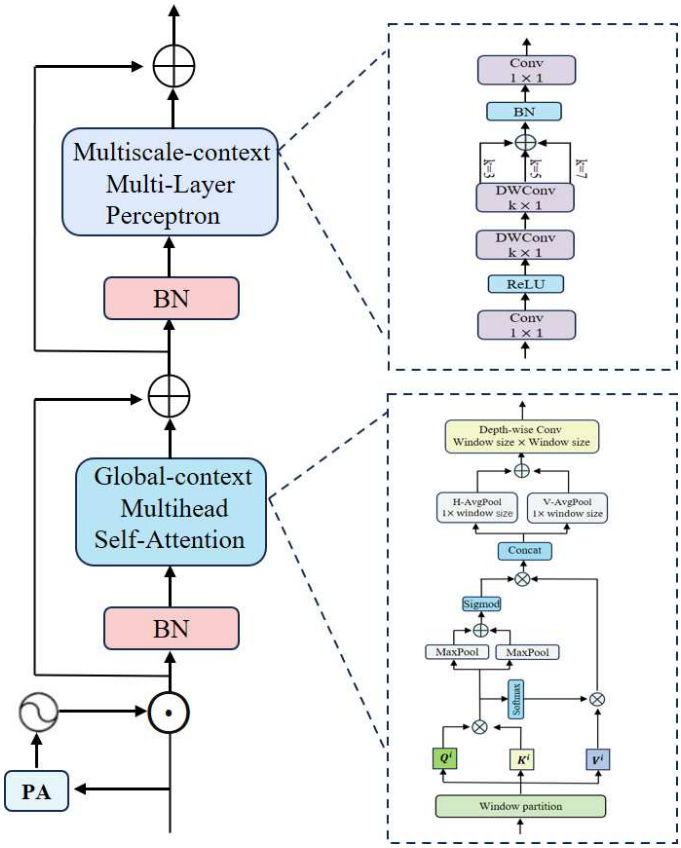


Fig. 3 The structure of the MGTB.

$$\text{head}^i = SA^i \odot CA^i \tag{4}$$

Each window's attention head is concatenated in position order. Then, the features undergo processing through horizontal and vertical window-based strip convolution, further enhancing the spatial information interaction between different regions. The processed features are fused by an addition operation and passed through a 3×3 depthwise separable convolution for further feature extraction, ultimately generating the output of the G-MCA module. The computation process can be expressed as:

$$X' = \text{Concat}(\text{head}^1, \text{head}^2, \ldots, \text{head}^n) \tag{5}$$

$$X_{\text{MSA}} = f_{\text{DWconv}}^{(3\times3)}\left(P_h^{(1\times S)}(X') \oplus P_V^{(S\times1)}(X')\right) \tag{6}$$

Where S is the window size, and $P_h^{1\times S}$ and $P_V^{S\times1}$ represent horizontal and vertical pooling operations in the window form, respectively.

For the output of G-MSA, we introduce a residual connection and feed it as the input to M-MLP. First, M-MLP applies a 1×1 convolution followed by a ReLU activation function to enhance the input features non-linearly. Then, it employs strip convolutions with three different kernel sizes (3, 5, 7) to extract multi-scale information and integrates detailed features across different scales through a summation operation.

Subsequently, batch normalization (BN) and a 1×1 convolution are applied to perform feature normalization and dimensional transformation. Finally, a residual connection is added to the output of M-MLP, which serves as the final output of MGTB. The following formulas can summarize the above process:

$$X_{\text{in}} = X_{\text{MSA}} + X \tag{7}$$

$$X_{\text{in}}^{'} = ReLU(f_{\text{conv}}^{1\times1}(X_{\text{in}})) \tag{8}$$

$$X_k = f_{\text{DWconv}}^{(1\times k)}\left(f_{\text{DWconv}}^{(k\times1)}(X_{\text{in}}^{'})\right) \tag{9}$$

$$X^* = \text{BN}\left(f_{\text{conv}}^{1\times1}(X_3 + X_5 + X_7)\right) \tag{10}$$

$$X_{\text{out}} = X^* + X_{\text{in}} \tag{11}$$

where $k = 3, 5, 7$ denote stripe convolution kernel sizes for capturing small, medium, and large-scale contextual information, and X_{out} represents the output features of MGTB.

3.4 Feature Fusion Module

High-level and low-level feature maps contain different levels of semantic information, and fusing them enables more effective perception of object scale variations. This multi-scale feature integration helps improve segmentation accuracy, as low-level features retain fine-grained local details, while high-level features provide rich global context. However, most existing methods typically fuse these features via simple element-wise addition after upsampling or channel-wise concatenation, which may lead to information imbalance and compromise model performance. To address this, we design a Feature Fusion Module (FFM) to more effectively combine the local details from low-level features and the global context from high-level features, as illustrated in Fig.4.

Specifically, the low-level feature F_l and high-level feature F_h are first processed by a 1×1 convolution and an upsampling operation, respectively, to generate F_l' and F_h' with the same resolution and number of channels. Then, two parallel branches are applied

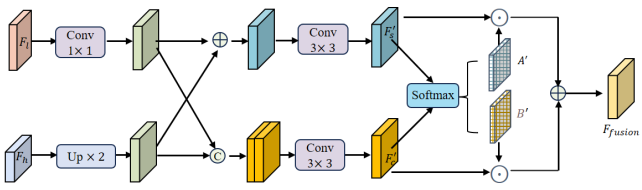


Fig. 4 Structure of the FFM module.

to these aligned features to obtain spatial and channel representations, denoted as F_s' and F_c', respectively. -In the spatial branch, we perform an element-wise sum of F_l' and F_h', followed by a 3×3 convolution to enhance important spatial context and suppress irrelevant information. In the channel branch, we concatenate F_l' and F_h' along the channel dimension and apply a 3×3 convolution to strengthen feature representation and restore the channel number. Finally, F_s' and F_c' are passed through a softmax layer to generate two attention maps A' and B', which adaptively weight and fuse the spatial and channel features. The overall process can be formulated as:

$$F_s' = f_{\text{conv}}^{3\times3}(F_l' + F_h') \tag{12}$$

$$F_c' = f_{\text{conv}}^{3\times3}(\text{Concat}(F_l', F_h')) \tag{13}$$

$$(F_s', F_c') \rightarrow \text{Softmax} \rightarrow (A', B') \tag{14}$$

$$F_{\text{fusion}} = A' \odot F_s' + B' \odot F_c' \tag{15}$$

where $\odot$ denotes element-wise multiplication, and F_{fusion} is the final fused feature output of FFM.

4 Experiments and results

4.1 Datasets

We conducted experiments on two well-known remote sensing datasets, Vaihingen and Potsdam, to evaluate the segmentation performance of our proposed MCANet.

- Vaihingen: The Vaihingen dataset, named after the Vaihingen area of Stuttgart, Germany, is a publicly available high-resolution aerial image dataset. It contains 33 high-resolution orthophoto images along with corresponding ground truth labels. Each image has an average size of 2494 × 2064 pixels. The ground truth labels annotate each pixel with its semantic class, such as impervious surfaces, buildings, low vegetation, trees, and vehicles. In this study, we selected 16 images for training, and the remaining images were used for testing. Each image is cut into small blocks of 256×256 to meet the experimental requirements.
- Potsdam: The Potsdam dataset consists of aerial images from the Potsdam area in Germany and is a commonly used public dataset in computer vision and remote sensing image processing. It contains high-resolution aerial images of various urban landscapes, covering multiple categories such as buildings, roads, and grass. Each image is accompanied by detailed ground truth annotations that label different types of ground objects. The Potsdam dataset includes 38 high-resolution images, each with a size of 6000 × 6000 pixels. In our experiments, 16 images were used for training, and the remaining ones were used for testing. Similar to the Vaihingen dataset, each image was also cropped into 256 × 256 to satisfy experimental needs.

4.2 Experimental Setup

Our network is built on the Pytorch framework, using the SGD optimizer with a momentum term of 0.9 and a weight decay of 1e-4 to train the model. Additionally, we set the initial learning rate to 0.01 and use the "poly"

Table 1 Quantitative comparison between our method (MCANet) and state-of-the-art methods on the Vaihingen dataset.

Method	F1%					mIoU%	mF1%	Param./M	FLOPS/G
	Imp.surf	Building	Low.veg	Tree	Car				
FCN [1]	85.86	89.34	75.62	83.72	75.33	69.84	81.79	19.17	25.54
DeeplabV3+[3]	87.74	91.55	76.80	85.01	73.99	71.52	83.02	59.34	65.27
DANet[27]	85.92	90.79	75.50	83.57	60.48	66.85	79.25	47.44	**13.85**
PSPNet [28]	84.33	88.03	74.50	83.33	58.38	64.71	77.71	52.52	50.51
MANet[29]	86.19	89.67	75.68	84.42	76.44	70.56	82.48	35.86	77.75
SwinUnet [30]	86.27	89.01	76.22	84.30	71.26	69.17	81.41	41.34	68.34
TransUnet [12]	87.75	91.74	77.21	84.56	76.86	72.29	83.62	100.44	35.84
UnetFormer[14]	87.54	91.34	76.43	84.77	77.42	72.10	83.50	**11.48**	16.99
CMLFormer[15]	88.01	91.90	77.26	84.87	78.82	73.07	84.18	11.81	21.25
MCANet(ours)	**88.25**	**92.04**	**77.51**	**85.24**	**79.13**	**73.54**	**84.43**	12.55	20.31

strategy for updating. All experiments are implemented on a single NVIDIA Tesla A100 40GB GPU, with a batch size of 8 and a maximum training epoch of 80. The overall loss for training the model is set as follows:

$$L_{\text{Total}} = L_{\text{CE}} + 0.5 \times L_{\text{Dice}} + 0.5 \times L_{\text{Focal}} \tag{16}$$

where L_{CE} is the cross-entropy loss, and L_{Dice} and L_{Focal} represent the Dice loss and Focal loss, respectively.

For evaluating the model performance, we use the mean F1 score (mF1) and mean intersection over union (mIoU) as evaluation metrics. These two evaluation metrics are based on the confusion matrix, which contains four elements: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The specific formulas are as follows:

$$\text{Precision} = \frac{1}{N}\sum_{j=1}^{N}\frac{TP_j}{TP_j + FP_j} \tag{17}$$

$$\text{Recall} = \frac{1}{N}\sum_{j=1}^{N}\frac{TP_j}{TP_j + FN_j} \tag{18}$$

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{19}$$

$$mIoU = \frac{1}{N}\sum_{j=1}^{N}\frac{TP_j}{TP_j + FP_j + FN_j} \tag{20}$$

4.3 Experimental results and Analysis

To verify the superiority of the proposed MCANet, this paper compares it with several mainstream methods on the Vaihingen and Potsdam remote sensing datasets, including classic CNN-based semantic segmentation methods such as FCN, Deeplabv3+, DANet, PSPNet, and MANet, Transformer-based methods like SwinUnet, and hybrid methods that integrate CNN and Transformer architectures, such as UnetFormer, TransUnet, and CMLFormer.

Results on the Vaihingen Dataset. Table 1 presents the results of a quantitative comparison between MCANet and other methods on the Vaihingen dataset. Specifically, MCANet achieves outstanding performance with mIoU and mF1 scores of 73.54% and 84.43%, respectively, surpassing all compared methods. Compared with the best-performing CNN-based method, DeeplabV3+, our proposed method improves mIoU and mF1 by 2.02% and 1.41%, respectively. Compared to the Transformer-based SwinUnet, MCANet achieves gains of 4.37% and 3.02% on these two metrics. Meanwhile, MCANet demonstrates significant advantages in parameter size and computational cost, with only 12.55M parameters and 20.31 GFLOPs, which are substantially lower than those of DeeplabV3+ and SwinUnet. In addition, when compared with other hybrid-architecture-based methods, MCANet also shows superior segmentation performance. Compared to the previously best-performing CMLFormer, MCANet still achieves a lead of 0.47% in mIoU and 0.25% in mF1.

Fig.5 illustrates the qualitative comparison results of MCANet and other methods on the Vaihingen dataset. In the first row, some methods exhibit misclassification or blurry boundaries for vehicles under shadow interference. In contrast, MCANet can effectively recognize vehicles and accurately delineate their boundaries, clearly outperforming the others. The second and third rows reveal semantic overlap issues due to the similar height of low vegetation and trees, making it challenging to distinguish between them. Conversely, MCANet presents smoother and clearer segmentation boundaries when handling these similar and overlapping categories, effectively reducing inter-class confusion. In the dashed box at the upper right corner of the fourth row, we highlight the incorrect identification of impervious surfaces by other methods under challenging lighting conditions and shadow occlusions. Nevertheless, MCANet consistently maintains accurate segmentation, demonstrating robustness and effectively mitigating the adverse impacts caused by lighting variations and shadows.

Results on the Potsdam Dataset. We report the quantitative comparison results on the Potsdam dataset in Table 2. The proposed MCANet achieves 80.30% mIoU and 89.24% mF1, representing the current best segmentation performance. Compared with the CNN-based MANet, MCANet improves mIoU and mF1 by 1.18% and 1.07%, respectively. Compared with the Transformer-based SwinUnet, MCANet outperforms by 1.62% and 1.35% on these two metrics, demonstrating its superiority in remote sensing image segmentation tasks. In addition, compared to CMLFormer, which also

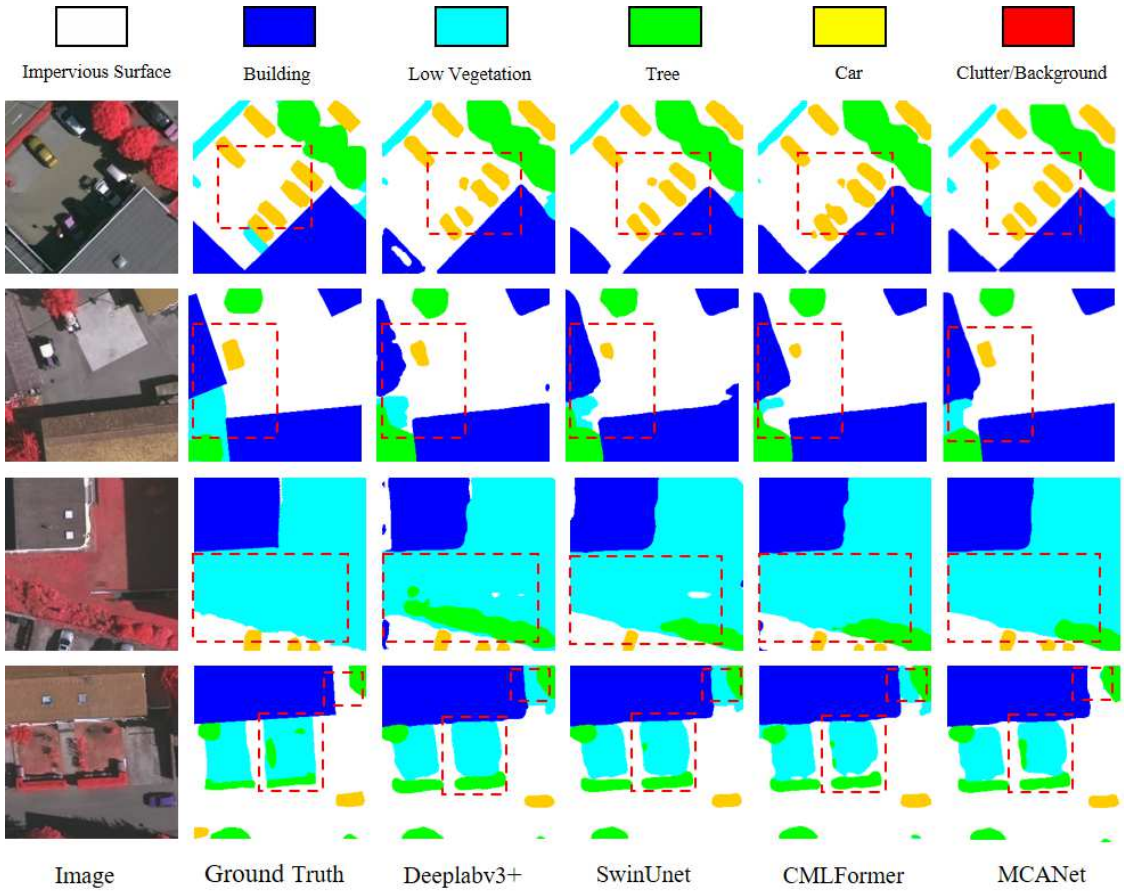


Fig. 5 Qualitative comparison between our MCANet and other methods on the Vaihingen dataset.

Table 2 Quantitative comparison between our method (MCANet) and state-of-the-art methods on the Potsdam dataset.

Method	F1%					mIoU%	mF1%	Param./M	FLOPS/G
	Imp.surf	Building	Low.veg	Tree	Car				
FCN[1]	89.76	93.64	82.95	83.12	89.82	78.60	87.86	19.17	25.54
DeeplabV3+[3]	90.40	95.01	83.82	84.22	87.32	79.07	88.15	59.34	65.27
DANet [27]	89.35	94.71	82.44	84.18	82.71	76.81	86.68	47.44	13.85
PSPNet [28]	89.45	93.38	82.13	80.00	89.36	77.12	86.86	52.52	50.51
MANet [29]	90.10	94.38	83.11	83.42	89.85	79.12	88.17	35.86	77.75
SwinUnet [30]	90.30	94.41	83.05	82.83	88.84	78.68	87.89	41.34	68.34
TransUnet [12]	90.52	94.68	83.87	83.58	90.24	79.76	88.58	100.44	35.84
UnetFormer [14]	90.14	94.44	83.15	83.16	90.07	79.16	88.19	11.48	16.99
CMLFormer [15]	90.79	94.96	84.01	83.91	90.31	80.06	88.79	11.81	21.25
MCANet(ours)	**90.96**	**95.42**	**84.78**	**84.51**	**90.53**	**80.30**	**89.24**	12.55	20.31

adopts a hybrid architecture, MCANet achieves 0.24% and 0.45% higher mIoU and mF1, respectively. We further provide per-class F1 scores, showing that MCANet surpasses other methods across all five semantic categories. Particularly for buildings, low vegetation, and trees, MCANet demonstrates more significant advantages.

We present the qualitative segmentation results of MCANet compared to other methods on the Potsdam dataset in Figure 6. In the first example, we observe that MCANet achieves much more precise and well-defined boundary segmentation for low vegetation within the dashed box compared to other methods. In the third row, we note that other methods mistakenly classify trees within the dashed box as low vegetation due to variations in shadow and illumination, leading to category confusion and blurred boundaries. In contrast, MCANet produces results that are more aligned with the ground truth, offering a clearer and more accurate distinction. In the second and fourth rows, we showcase more complex scenes with buildings or low vegetation next to cluttered backgrounds. It is evident that other methods frequently struggle with blurry boundaries or missing regions, while MCANet excels in accurately differentiating targets from the background, generating smoother and more precise segmentation boundaries. This further highlights MCANet's exceptional performance in fine-grained segmentation tasks, demonstrating its ability to preserve intricate details in challenging scenes.

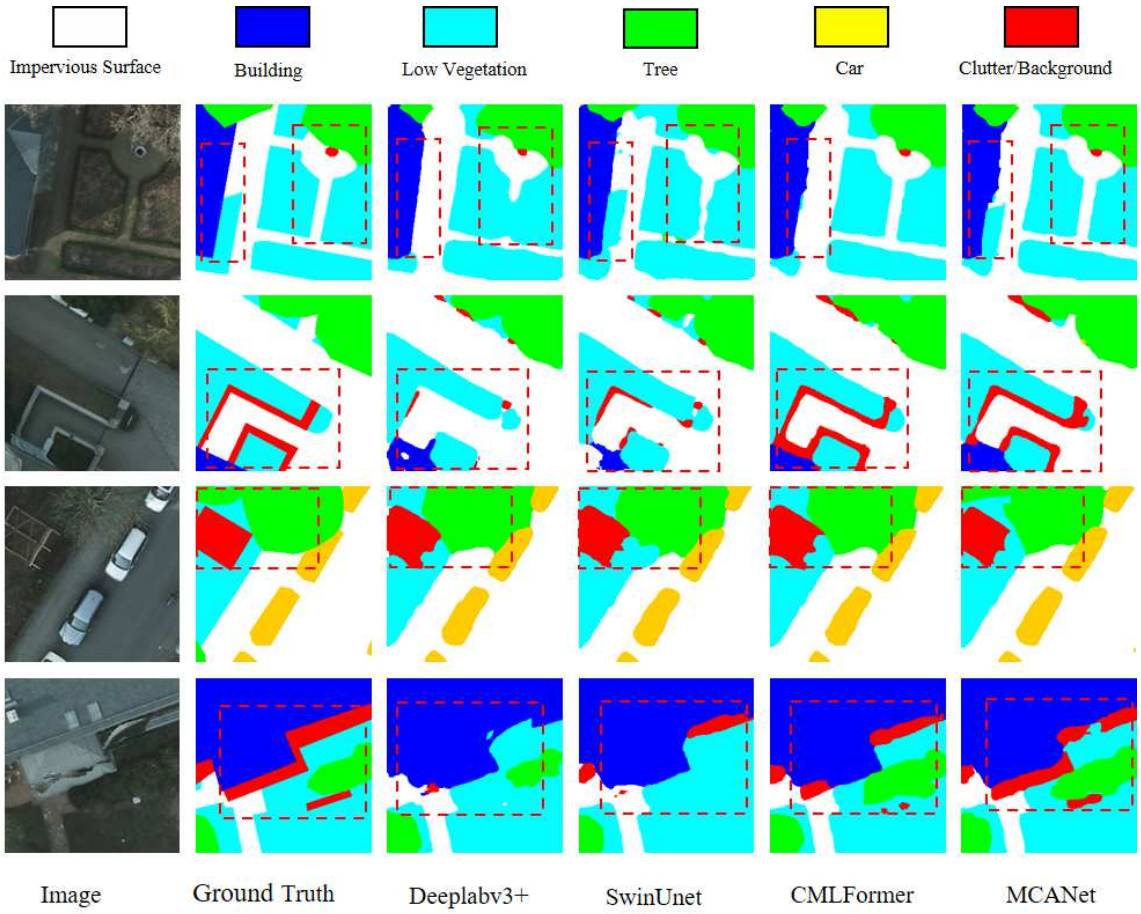


Fig. 6 Qualitative comparison between our MCANet and other methods on the Potsdam dataset.

Table 3 Validation of the effectiveness of MCANet components on Vaihingen dataset.

Method	Aggregation	F1%					mIoU%	mF1%	Param./M
	S&C	Imp.surf	Building	Low.veg.	Tree	Car			
ResNet-18(baseline)	✗	86.13	90.21	74.96	83.65	65.18	67.61	80.05	11.42
baseline + MGTB	✗	87.42	91.17	76.54	84.87	72.36	70.16	82.47	12.15
baseline + MGTB + AFEM	✗	87.67	91.51	77.11	84.96	77.79	71.94	83.81	12.39
baseline + MGTB + AFEM + FFM	✗	88.07	91.85	77.33	85.12	78.64	73.19	84.20	12.50
baseline + MGTB + AFEM + FFM	✓	**88.25**	**92.04**	**77.51**	**85.24**	**79.13**	**73.54**	**84.43**	12.55

4.4 Ablation study

To verify the effectiveness of the components in MCANet, systematic ablation experiments were conducted on the Vaihingen dataset, and the results are shown in Table 3. Since a lightweight ResNet-18 was selected as the backbone network, we chose to use the output results by connecting the features from different stages of ResNet-18 through upsampling as the baseline for comparison. Before introducing FFM, we used the skip connection approach from U-Net to fuse the encoder and decoder features. In the experiment, mF1 and mIoU were used as evaluation metrics, while the number of parameters measured the model's lightweight nature. The first row shows the segmentation performance and parameter count of the baseline model on the Vaihingen dataset. The following rows display the experimental results after introducing the various components of MCANet. When MGTB was introduced on top of the baseline model, mF1 and mIoU improved by 2.55% and 2.42%, respectively. The introduction of MGTB significantly improved the F1 scores of each category, which can be attributed to the multi-scale global context information captured by MGTB. By effectively integrating global features from different scales, MGTB could better recognize the details and overall structure in the images, thereby improving segmentation accuracy. The AFEM module enhanced the local context information adaptively, effectively improving the feature representation ability, resulting in a noticeable improvement in segmentation accuracy, with mIoU and mF1 reaching 71.94% and 83.81%, respectively, validating the critical role of the AFEM module in capturing local detail information and optimizing target boundaries. Finally, with the addition of the FFM module, the model's mIoU and mF1 were further improved by 1.25% and 0.39%, respectively. Notably, in our network architecture, the features from the three stages of the decoder, after upsampling and summing, are concatenated with

the outermost features, and the final predicted output map is generated through the segmentation head. Here, we use S&C to represent this feature aggregation operation method, and by comparing the experimental results before and after its introduction, we validate the effectiveness of this approach. The results show that after introducing the S&C operation, the model's mIoU and mF1 further improved by 0.33% and 0.23%, achieving the best performance of 73.54% mIoU and 84.43% mF1. This demonstrates that upsampling, summing, and concatenating features from different scales can effectively enhance feature representation ability, thereby improving segmentation accuracy.

5 Conclusion

In this paper, we propose MCANet, a novel remote sensing image semantic segmentation method that combines CNN and Transformer architectures. To leverage both global and local contextual information, we design the AFEM module to enhance local features from the encoder. The MGTB module in the decoder captures long-range dependencies using multi-head self-attention and multi-scale strip convolution to extract global context. Additionally, we introduce the FFM module to fuse local and global features, improving segmentation performance. Extensive comparison and ablation experiments on the Vaihingen and Potsdam datasets validate the effectiveness of MCANet. Furthermore, MCANet optimizes computational efficiency while maintaining high segmentation accuracy, making it suitable for real-world applications. In the future, our work will focus on further optimizing the model and exploring its application to larger datasets and other remote sensing tasks.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62162007) and the Guizhou Provincial Basic Research Program (Natural Science) (No. QianKeHeJiChu-ZK[2024]YiBan079).

Declarations

- **Competing Interest** All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- **Ethical Declaration** This article does not contain any studies with human participants or animals performed by any of the authors.
- **Data availability** The data relevant to this research is available at https://www.isprs.org/.

References

[1] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015). https://doi.org/10.1109/TPAMI.2016.2572683

[2] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28 . Springer

[3] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018). https://doi.org/10.1007/978-3-030-01234-2_49

[4] Cao, R., Ning, L., Zhou, C., Wei, P., Ding, Y., Tan, D., Zheng, C.: Cfanet: Context feature fusion and attention mechanism based network for small target segmentation in medical images. Sensors **23**(21), 8739 (2023) https://doi.org/10.3390/s23218739

[5] Long, W., Zhang, Y., Cui, Z., Xu, Y., Zhang, X.: Threshold attention network for semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–12 (2023) https://doi.org/10.1109/TGRS.2023.3276081

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[7] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., *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, pp. 6881–6890 (2021). https://doi.org/10.1109/CVPR46437.2021.00681

[8] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986

[9] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021). https://doi.org/10.1109/ICCV48922.2021.00061

[10] Zhang, Q., Yang, Y.-B.: Rest: An efficient transformer for visual recognition. Advances in neural information processing systems **34**, 15475–15485 (2021)

[11] Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021)

[12] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[13] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y.: Swin transformer embedding unet for remote sensing image semantic segmentation. IEEE transactions on geoscience and remote sensing **60**, 1–15 (2022) https://doi.org/10.1109/TGRS.2022.3144165

[14] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M.: Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 196–214 (2022)

https://doi.org/10.1016/j.isprsjprs.2022.06.008
[15] Wu, H., Zhang, M., Huang, P., Tang, W.: Cmlformer: Cnn and multi-scale local-context transformer network for remote sensing images semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024) https://doi.org/10.1109/JSTARS.2024.3375313
[16] Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., Wang, C.: Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–20 (2022) https://doi.org/10.1109/TGRS.2022.3144894
[17] Wang, Q., Jin, X., Jiang, Q., Wu, L., Zhang, Y., Zhou, W.: Dbct-net: A dual branch hybrid cnn-transformer network for remote sensing image fusion. Expert Systems with Applications **233**, 120829 (2023) https://doi.org/10.1016/j.eswa.2023.120829
[18] Gao, L., Liu, H., Yang, M., Chen, L., Wan, Y., Xiao, Z., Qian, Y.: Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. IEEE journal of selected topics in applied earth observations and remote sensing **14**, 10990–11003 (2021) https://doi.org/10.1109/JSTARS.2021.3119654
[19] Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175–12185 (2022). https://doi.org/10.1109/CVPR52688.2022.01186
[20] Wang, Z., Zhang, S., Zhang, C., Wang, B.: Hidden feature-guided semantic segmentation network for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–17 (2023) https://doi.org/10.1109/TGRS.2023.3244273
[21] Li, R., Duan, C., Zheng, S., Zhang, C., Atkinson, P.M.: Macu-net for semantic segmentation of fine-resolution remotely sensed images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022) https://doi.org/10.1109/LGRS.2021.3052886
[22] Chen, K., Xia, M., Lin, H., Qian, M.: Multiscale attention feature aggregation network for cloud and cloud shadow segmentation. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–16 (2023) https://doi.org/10.1109/TGRS.2023.3283435
[23] Peng, C., Zhang, K., Ma, Y., Ma, J.: Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021) https://doi.org/10.1109/TGRS.2021.3053062
[24] Hang, R., Yang, P., Zhou, F., Liu, Q.: Multiscale progressive segmentation network for high-resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–12 (2022) https://doi.org/10.1109/TGRS.2022.3207551
[25] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
[26] Polesel, A., Ramponi, G., Mathews, V.J.: Image enhancement via adaptive unsharp masking. IEEE transactions on image processing **9**(3), 505–510 (2000) https://doi.org/10.1109/83.826787
[27] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019). https://doi.org/10.1109/CVPR.2019.00326
[28] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017). https://doi.org/10.1109/CVPR.2017.660
[29] Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M.: Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021) https://doi.org/10.1109/TGRS.2021.3093977
[30] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). https://doi.org/10.1007/978-3-031-25066-8_9