AANet: Adaptive Attention Networks for Semantic Segmentation of High-Resolution Remote Sensing Imagery

Yan Chen, Qianchuan Zhang, Xiaofeng Wang, Quan Dong, Menglei Kang, Wenxiang Jiang, Mengyuan Wang, Lixiang Xu, and Chen Zhang

***Abstract*—Contextual information can effectively aid deep-learning models in extracting interclass and intraclass difference features in remote sensing images. This article presents a novel approach called the adaptive attention network (AANet) for semantic segmentation in high-resolution remote sensing images. The proposed AANet aims to enhance the segmentation performance while minimizing the network's computational and parametric aspects. Furthermore, the AANet is designed to facilitate real-time segmentation. The AANet involves the construction of three distinct modules, namely the multiscale channel attention module (MCAM), the multidimensional spatial attention module (MSAM), and the contextual information adaptive fusion module (CIAFM). MCAM enhances a multiscale approach to effectively capture contextual information from neighboring channels and category information. MSAM is designed to extract and combine detailed information from each dimension of the spatial domain. CIAFM focuses on the complementary nature of channel and spatial context information and the correlation between pixels and categories. The methodology employed in this article involved conducting experiments on the ISPRS Vaihingen, ISPRS Potsdam, and multiobject coastal supervision semantic segmentation dataset (MO-CSSSD) datasets alongside a comparative analysis with conventional semantic segmentation models. The results of the article indicate that our approach demonstrates exceptional performance on the ISPRS Vaihingen dataset, ISPRS Potsdam dataset, and MO-CSSSD dataset, achieving mean intersection over union scores of 83.17%, 85.67%, and 89.68%, respectively.**

***Index Terms*—Adaptive attention, contextual information, high-resolution remote sensing imagery, multidimensional spatial attention, multiscale channel attention.**

Manuscript received 6 March 2024; revised 2 May 2024 and 19 June 2024; accepted 7 August 2024. Date of publication 19 August 2024; date of current version 26 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176085; in part by the Key Scientific Research Foundation of the Education Department of Province Anhui under Grant KJ2020A0658; in part by the University Natural Sciences Research Project of Province Anhui under Grant KJ2021ZD0118; in part by the Hefei University Talent Research Funding under Grant 20RC13; in part by the Hefei University Scientific Research Development Funding under Grant 20ZR03ZDA; in part by the Program for Scientific Research Innovation Team in Colleges and Universities of Anhui Province under Grant 2022AH010095; and in part by the Hefei Specially Recruited Foreign Expert support. *(Corresponding author: Qianchuan Zhang.)*

The authors are with the School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China (e-mail: chenyan@hfuu.edu.cn; qianczhang@163.com; xfwang@hfuu.edu.cn; dq2112774778@163.com; kangml@stu.hfuu.edu.cn; jiangwx@stu.hfuu.edu.cn; wangmy@stu.hfuu.edu.cn; xulixiang@hfuu.edu.cn; zhangchen@hfuu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3443283

I. Introduction

SEMANTIC segmentation of remote sensing images, which aims at assigning a specific label to each pixel in the image, has become one of the most essential methods for the intelligent interpretation of ground information and plays a crucial role in several application scenarios, such as land cover mapping [1], [2], [3], building extraction [4], [5], [6], environmental protection [7], [8], and economic assessment [9], [10].

In recent years, with the rapid development of deep learning, the convolutional neural network (CNN) has demonstrated powerful hierarchical representation abilities, thanks to their characteristics such as local perception, parameter sharing, and multilayered structure. As a result, CNN has become the mainstream technology in the field of semantic segmentation. Fully convolutional neural network (FCN) [11] proved the first effective end-to-end CNN-based semantic segmentation method and laid the foundation for fully convolutional semantic segmentation. Although FCN achieved satisfactory results, the oversimplification of its decoder design hindered its ability to provide more comprehensive contextual information for semantic segmentation tasks in remote sensing imagery. Many improved methods for extracting global context information have been proposed on CNN-based architectures. For instance, UNet [12] is constructed using contraction paths to extract features and expansion paths to integrate high-dimensional and low-dimensional features to regain the native resolution. It provides more contextual information for image segmentation through skip connections. DeepLabV3 [13] is a CNN-based model that employs atrous convolution to capture global context information by expanding the size of the receptive field of the convolution. PSPNet [14] utilizes the technique of pyramid pooling to extract global information from images. These methods can capture local contextual information but offer limited insights into a global context. The global and local context information is delineated in Fig. 1. The variability and similarities of objects in remote sensing images present a challenge for their processing, dealing with minor differences between classes but large ones within each. The undertaking of semantic segmentation for remote sensing images presents a significant challenge. By incorporating the global context information of an image into the semantic segmentation task for remote sensing images, the model gains a more comprehensive understanding of the interclass and intraclass difference features,

Fig. 1. Description of global vs. local contextual information. Local context information is learned by modeling with convolution operations (yellow) and global context information is learned by establishing remote window dependencies (red).

thereby assisting in addressing the challenges encountered when processing remote sensing images.

Recently, attention mechanisms have brought new ideas to extract global contextual information. Channel attention can provide more accurate semantic category information. SENet [15] uses a global average pooling operation to compress all channel information into scalar features. These nonlinear features are subsequently processed using a multilayer perceptron. The activation function Sigmoid is utilized to recalibrate the feature information, which enhances significant areas while diminishing nonimportant spots. The FcaNet [16] smartly combines the discrete cosine transform [17] with the channel attention mechanism to improve the squeeze module in SENet. When considering the same network complexity, the segmentation accuracy exhibits a 1.8% improvement compared to SENet. Spatial attention is an essential technology in deep CNNs to extract spatial context information. It strengthens the network's ability to obtain representative features in the feature map by screening critical areas. For example, CCNet [18] uses the relationship between space and channels further to extend the attention mechanism to the cross-channel dimension, enhances the CNN's ability to interact with images across channels and regions, and captures contextual information between diverse channels, scales, and directions through cross paths. Although these attention models have achieved significant progress, they often introduce a lot of redundant noise when extracting global contextual information across channels, or they may ignore relevant information in different dimensions when extracting spatial contextual information. As a result, the model's ability to segment object categories is limited.

The complementarity and interaction of channel and spatial context information are exceptionally critical for semantic segmentation. Therefore, rational utilization of the related information can significantly enhance the segmentation performance of the network. In response to this supposition, many researchers have proposed hybrid attention mechanism networks. For example, Jiang et al. [19] added a spatial attention module and a channel attention module at the end of the UNet architecture. These two modules obtain the spatial and channel contextual information in a parallel manner and can better extract the feature information of interest. DANet [20] combines the channelwise contextual information and spatial contextual information extracted from two pathways by weighting them together, leading to promising results. These hybrid attention mechanism networks fuse spatial and channel context information through weighted summation or multiplication, but they do not fully exploit the complementary nature of channel and spatial context information, thereby impacting the network's performance in semantic segmentation tasks.

In recent years, the self-attention mechanism, which is of great interest in natural language processing, has been introduced into computer vision and achieved good results, especially in semantic segmentation. As an illustration, the module proposed by [21] encompasses a global self-attention mechanism and a local window self-attention mechanism. This design enables the module to effectively capture long-range semantic information and local details. NLNet [22] adopts a nonlocal attention mechanism to learn the relationships between pixels, thereby providing global contextual information for semantic segmentation tasks. Because networks based on self-attention mechanisms process images pixel by pixel, this would lead to significant resource consumption. Therefore, standard self-attention mechanisms (SSAMs) are not suitable for high-resolution image semantic segmentation applications.

In order to tackle the shortcomings of the approaches discussed above and to cope with the obstacles of processing high-resolution images, this article suggests an adaptive attention network (AANet) for the semantic segmentation of high-resolution remote sensing imagery. The AANet network utilizes an encoder–decoder architecture. The encoder part consists of ResNet [23], a lightweight CNN-based network, and the decoder part consists of several improved modules, including our proposed Multiscale channel attention module (MCAM), multidimensional spatial attention module (MSAM), and contextual information adaptive fusion module (CIAFM). Among them, MACM utilizes a multiscale strategy to extract adjacent

contextual and category information in the channel to enhance the discriminative ability of pixel categories in remote sensing image segmentation. MSAM performs attentional weighting in different dimensions to extract spatial contextual information in critical regions. CIAFM uses an improved self-attention mechanism to deeply fuse the contextual information of the channel with the space and utilize the category information to guide pixels for classification. This approach allows the model to further mine interclass and intraclass features in images while reducing the amount of computation and the number of parameters. Our proposed innovative modules and network are comprehensively evaluated experimentally on the widely used ISPRS Vaihingen and Potsdam datasets. The experimental findings demonstrate that AANet exhibits notable benefits in comparison to other popular lightweight networks for the purpose of remote sensing imagery segmentation tasks.

The main contributions of this article are summarized as follows.

1) Two improved parallel attention modules (MCAM and MSAM) oriented to channel and pixel space have been designed based on adaptive strategies to effectively fuse the feature information from multiple scales and dimensions, and minimize redundancies and suppress the influence of nonobject noise.
2) An enhanced lightweight self-attention mechanism (CIAFM) has been constructed to deeply integrate spatial and channel contextual information generated by proposed modules for the key information screening between categories and pixels, and to improve the segmentation ability of the model and optimize the computational complexity.
3) We proposed a novel AANet for semantic segmentation in high-resolution remote sensing images to ascend the segmentation performance while minimizing the network's computational and parametric aspects for facilitating real-time segmentation.

II. Related Work

A. Multiscale Semantic Segmentation Networks

In the semantic segmentation task of remote sensing images, considering the significant differences in the sizes of ground objects, adopting the multiscale strategy can effectively extract the feature information of target objects in the image, especially for small and large targets, thereby significantly improving the model's performance. Deeplabv3+ [24] has achieved tremendous success on public datasets by using parallel convolution operations with different dilation rates to extract multiscale feature information from the image and connect the obtained multiscale feature information. PSPNet [14] uses a pooling pyramid to fuse the hierarchically extracted multiscale feature information; this design captures the features of objects with varying sizes in space. It enables rich semantic details to interact with rich spatial information. Nie et al. [25] introduced a cross-scale interaction module to extract semantic features and uses convolutions with different dilation rates to extract multiscale boundary information. To capture more comprehensive multiscale details, various methods combined with attention mechanisms have been proposed. For instance, Liu and Lin [26] combined the SENet module, which extracts channel context information, with parallel convolutions of different receptive fields to capture multiscale features in the image, demonstrating significant effects on remote sensing datasets. Although networks incorporating channel attention mechanisms like SENet can effectively extract features, the dimensionality reduction operation leads to the loss of detailed channel information. ResUNet-a [27], using a U-Net backbone architecture, leverages multiple parallel residual atrous convolutions and spatial pyramid components to extract multiscale and contextual information from remote sensing images. In the HRNet-based [28] HRCNet [29] network, the LCA module extracts channel correlations, while the LSA module models and learns spatial information from the images. These extracted features are then combined using a weighted addition operation. Additionally, the FEFP module, which incorporates techniques, such as FPN and ASPP, is employed to learn multiscale features and contextual information from the images. PICS [30], which employs DeepLabV3+ with a ResNet101 backbone as its segmentation network, captures information at multiple scales using several parallel convolutions with different dilation rates. It also integrates multiple semisupervised paradigms to generate high-quality pseudolabeled samples. Additionally, PICS introduces a novel loss-based sample evaluation and selection method, further reducing the potential risk of error accumulation due to inevitable misclassifications. Zhang et al. [31] has designed two dual-attention modules with multiscale spatial attention and channel attention to extract multiscale feature information, contributing to target object recognition. Based on the transformer [32] architecture, Xiao et al. [33] has integrated a self-attention mechanism with an automatically adjustable receptive field, enabling feature extraction from target objects and capturing long-distance dependencies simultaneously. Although introducing transformer technology into multiscale networks enhances the model's ability to extract features, it also adds computational pressure to the model.

In this article, context information of neighboring channels is extracted by convolutions with different receptive fields. The extracted information is then weighted and summed to obtain the final feature representation. This approach not only mitigates the impact of channel dimension reduction but also reduces the complexity of the network.

B. Global Contextual Information Modeling

The complex diversity and similarity of features in remote sensing images lead to interclass and significant intraclass differences. Therefore, for the semantic segmentation task of remote sensing images, relying solely on local context information makes it difficult to accurately predict the category of each pixel. If global relations are introduced, this task will become much more straightforward. As a result, researchers have proposed various methods to extract global features, and among these methods, a common approach is to incorporate attention mechanisms into network architectures. For example, Guo et al. [34] adopted a network architecture composed of

multiple adaptive global average pooling layers with different size scales to extract rich spatial information. ECANet [35] has demonstrated the significant impact of channel dimensionality reduction on attention mechanism networks. Extracting adjacent contextual information through interchannel interactions not only enhance network performance but also significantly reduces the network's parameter and computational overhead. Li et al. [36] introduced a memory-efficient and computationally efficient linear attention mechanism into the skip connections of the UNet network architecture to extract global context information. Li et al. [37] proposed a novel kernel attention mechanism that reduces computational requirements and effectively extracts and utilizes global dependencies. In these attention networks that extract global contextual information, some have not noticed the importance of local knowledge when extracting global relations. At the same time, others have resulted in a large amount of redundant information when extracting channel information. Therefore, the application of these networks in high-resolution image segmentation is limited.

Researchers have also proposed some hybrid architecture methods to address the issues above. CBAM [38] concatenates the channel and spatial attention modules to form a hybrid attention module, which effectively extracts spatial and channel context. HMANet [39] introduces the CAA, CCA, and RSA attention modules. The CCA, combined with the CAA, is used to extract channel and category information, adaptively integrating the category information into the channel features. The RSA module learns spatial pixelwise correlations through regional shuffling. Finally, the extracted channel and spatial information are concatenated to produce the output features. Inspired by the self-attention mechanism, DMNet [40] introduces the PCRM and CCRM modules, which are designed to capture bidirectional semantic associations in the spatial and channel dimensions, respectively. These modules suppress category information specific to the support image itself, retaining more generalized common category semantic information. Additionally, the CSRM module is designed to extract specific semantic information to aid image segmentation, effectively addressing the issue of large intraclass variance. Fu et al. [20] used two independent attention branches to extract spatial and channel features. These are then weighted and summed to obtain global dependencies. Wang et al. [41] used self-attention and a CNN-based unit to extract global context and local detail information from remote sensing images. Finally, the extracted global and local information is processed through a feature refinement module to obtain useful results. Li et al. [42] utilized a spatial pathway to preserve spatial details and a context pathway to extract global contextual information from the image. Finally, a fusion module is developed to deeply integrate the spatial detail and context information. Long et al. [43] employed a dual encoder consisting of a self-attention-based encoder and a lightweight CNN-based encoder to extract global and local context. The self-attention-based encoder uses a sliding window mechanism to capture global context, while the CNN-based encoder focuses on local context.

Despite having so many advantages, converting each pixel into a sequence to compute dependencies between positions can be computationally expensive compared to convolutional operations, and it needs to take into account the complementarity between different dimensions in space. This article proposes an AANet for semantic segmentation in high-resolution remote sensing images. AANet utilizes a lightweight ResNet18 as the encoder and employs multiple modules in the decoder. The MCAM extracts adjacent channel details using an improved multiscale strategy. The MSAM pools feature and adaptively fuse information across different dimensions in space. The CIAFM employs a lightweight self-attention mechanism to deeply integrate channel and spatial context. Additionally, category information is utilized to guide pixel classification.

III. Methodology

A. Network Architecture

By increasing the attention on key regions, contextual information can effectively address the issue of slight interclass variance and significant intraclass variance caused by the abundance of detailed features in remote sensing images. This article proposes an AANet for semantic segmentation of high-resolution remote sensing images, as illustrated in Fig. 2. The network we proposed adopts an encoder–decoder architecture to fully utilize the hierarchical information extracted by the encoder. Many semantic segmentation methods for remote sensing images [44], [45], [46], [47], [48] use ResNet as the encoder of the network because it addresses the issues of gradient vanishing and model degradation during the training of deep neural networks by introducing residual blocks.

To ensure the efficiency of network inference, the proposed method in this article selects the lightweight version of ResNet, ResNet18, as the encoder of the network. This choice is made because ResNet18, compared to other deeper ResNet models, has lower computational complexity while maintaining model performance. The decoder of the network proposed in this article consists of multiple stages, each containing three novel modules: the MCAM for extracting channel contextual information and category information, the MSAM for enhancing attention on critical regions, and the CIAFM for exploring the complementarity between channels and spatial dimensions.

B. Multiscale Channel Attention Module (MCAM)

Different channel feature maps correspond to diverse semantic response information, and interdependencies exist between adjacent channels. Utilizing the correlation information between channels can enhance the representation of semantics, thereby achieving better segmentation results. As shown in Fig. 3, we propose the MCAM. MCAM utilizes a multiscale strategy to address the issue of significant size differences among objects in remote sensing images. Specifically, MCAM employs convolutional operations with different receptive field sizes to extract global information from adjacent channels. This approach helps the model perceive target features of different scales. Additionally, MCAM extracts category information between channels through convolutional and normalization operations. This can be used to guide pixel classification, further enhancing the model's ability to explore interclass and intraclass differing features.

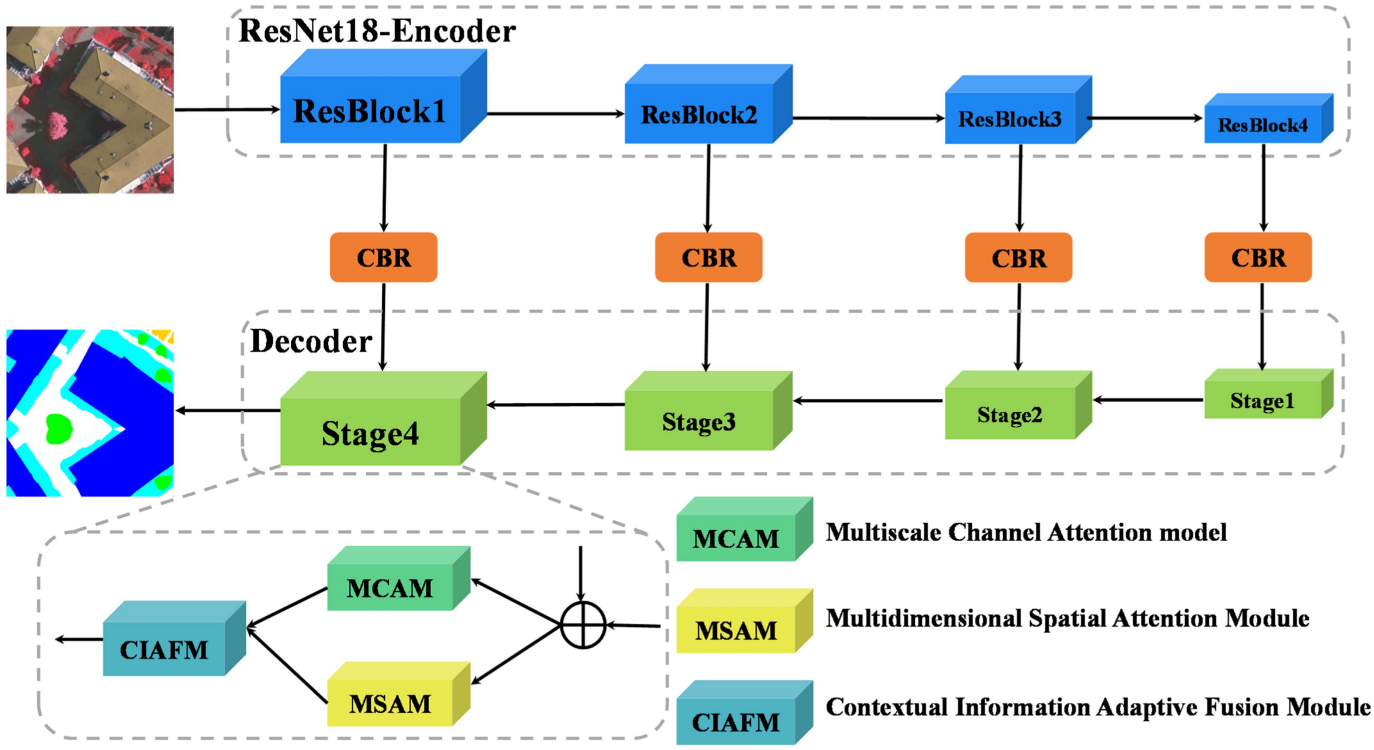


Fig. 2. Overview of the AANet.

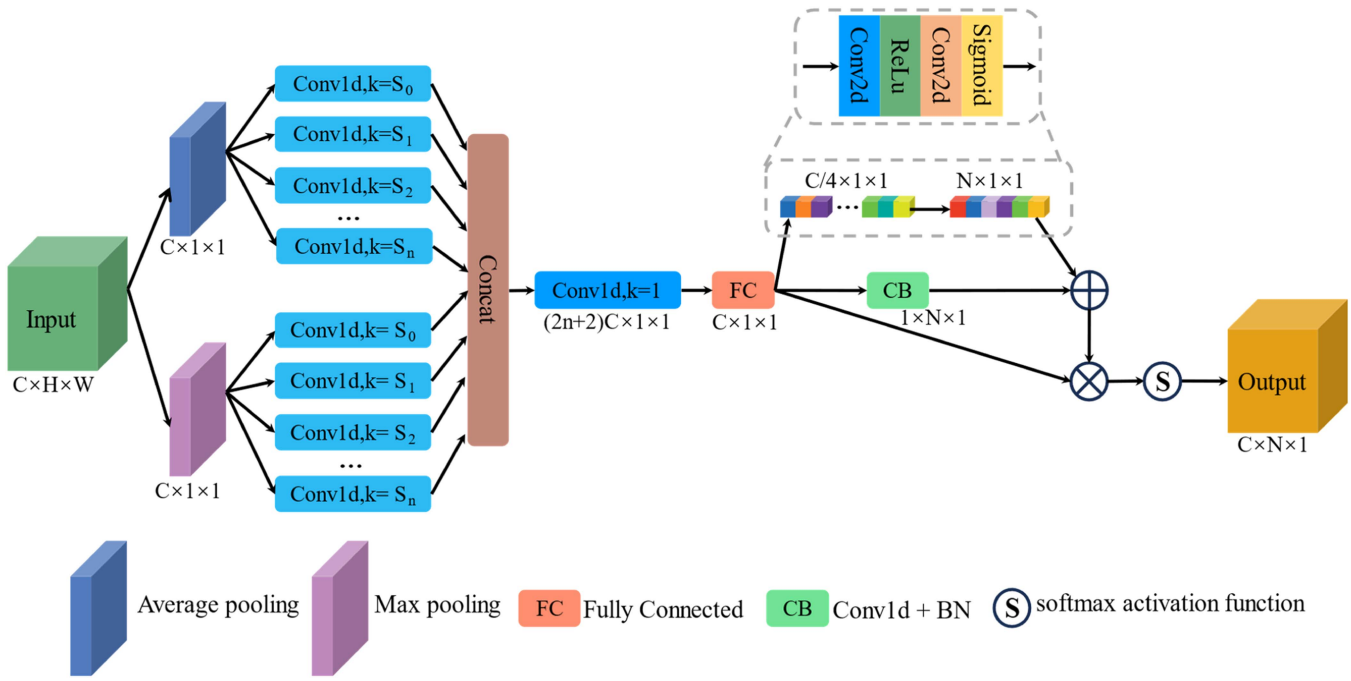


Fig. 3. Multiscale channel attention module.

We apply both average pooling and max pooling operations with a global receptive field to the feature map $X \in R^{C\times H\times W}$, resulting in $X_a \in R^{C\times 1\times 1}$ and $X_b \in R^{C\times 1\times 1}$. The average pooling operation and the max pooling operation can be represented as follows:

$$X_a = \text{avgpool}\,(X) \tag{1}$$

$$X_b = \text{maxpool}\,(X) \tag{2}$$

where avgpool is average pooling and maxpool is max pooling. ECANet [31] describes the relationship between the size of the convolution kernel and the number of channels, and its calculation process is shown in (3). A single-scale approach is used in ECANet to extract contextual information in the

channel, making it challenging to capture target features with large size differences in remotely sensed imagery. Therefore, this article has improved the single-scale convolution in ECANet by designing a multiscale module that replaces the single receptive field convolution with multiple parallel convolutions of different receptive fields. This aims to extract contextual information from adjacent channels at various scales, facilitating the perception of target features at different sizes in the image. Due to the wide range of sizes for target objects, we choose to use convolution with exponentially increasing kernel sizes to achieve multiscale perception of target objects at different levels in the image and to learn more abstract features during the training process. The exponentially increasing kernel sizes are calculated as follows:

$$S_0 = \left| \frac{(\log_2 C + 1)}{2} \right|_{\text{odd}} \tag{3}$$

$$S_i = 2^{i+\log_2(S_0-1)} + 1, (i = 1, 2, 3 \ldots n) \tag{4}$$

where C represents the number of channels, S_0 denotes the initial convolution kernel size, S_i denotes the size of the ith convolution kernel after incremental calculation (where $S_i \leq C$), and "odd" indicates selecting the nearest odd value.

In order to capture the multiscale features in the channel, we compute the coarse-grained global feature information X_a and X_b to obtain $X_c \in R^{8C\times1\times1}$ as follows:

$$X_c = \text{Concat}\left(\text{conv1d}(X_a)_i, \text{conv1d}(X_b)_i\right), (i=1,2,3,4) \tag{5}$$

where Concat is the splicing operation and $\text{conv1d}(X)_i$ represents the ith one-dimensional (1-D) convolution operation. To further smooth the local feature information, we recalibrate the aggregated features using a 1-D convolution with a kernel size of 1 and then extract the final neighboring channel contextual information $X_d \in R^{C\times1\times1}$ from the recalibrated global feature using fully connected layers. The complementary nature of categories and spatial pixels can effectively assist in image segmentation. To obtain N category features from the channels, we employ two sets of 2-D convolution operations and use 1-D to extract category information from the channel. Subsequently, we add the weights of the two types of category details, perform a transpose operation, and multiply the transposed category result with the local channel context to obtain the output of MCAM.

MCAM can provide channel context and N category information, which can be deeply integrated with spatial features, thereby establishing a profound correlation between channels and space, as well as between pixels and categories.

C. Multidimensional Spatial Attention Module (MSAM)

In remote sensing images, objects often have similar shapes and contours due to their shared environment. Addressing this high similarity issue relying on local spatial information is challenging. Incorporating spatial global dependencies can effectively resolve the problem of high spatial similarity. Consequently, this article proposes the MSAM for extracting and integrating global contextual information across various dimensions in space. The structure of MSAM is illustrated in Fig. 4.

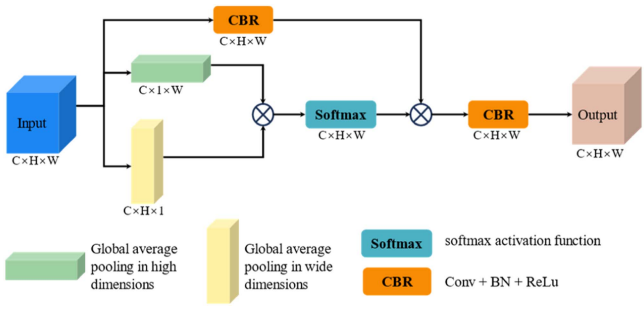


Fig. 4. Multidimensional spatial attention module.

To obtain feature information in different dimensions of space, we use two global pooling operations of various sizes on the input feature map $X \in R^{C\times H\times W}$ to extract the vertical and horizontal dimension information in space, resulting in feature maps $X_H \in R^{C\times H\times1}$ and $X_W \in R^{C\times1\times W}$, respectively. Our vertical pooling and horizontal pooling can be described as follows:

$$X_H = V(X) \tag{6}$$

$$X_W = H(X) \tag{7}$$

where V is the vertical global average pooling. H represents horizontal global average pooling. To adaptively integrate feature information across different dimensions in space and make the pixel information in space easier to interpret, we execute a coarse feature fusion for each dimension as follows:

$$X_{HW} = \text{reshape}(X_H) \otimes \text{reshape}(X_W) \tag{8}$$

where reshape represents size reshaping of the feature map and $\otimes$ denotes matrix multiplication. Finally, we use the activation function and convolution operation with a kernel size of 3 to obtain fine-grained spatial feature information $X_{\text{SA}} \in R^{C\times H\times W}$, and this step can be described as

$$X_{\text{SA}} = \text{conv}(\sigma(X_{HW}) \otimes X) \tag{9}$$

where $\otimes$ represents the feature map multiplication, σ represents the activation function, and conv denotes the 2-D convolution with a kernel size of 3. MSAM extracts the feature information of different dimensions in the space and does the adaptive fusion, which increases the model's focus on the key regions and suppresses the interference of useless features.

D. Contextual Information Adaptive Fusion Module (CIAFM)

Previous article has demonstrated that the relationship between channels and spatial features, as well as between categories and pixels, can significantly improve the segmentation performance of remote sensing images. As shown in Fig. 5, this article proposes the CIAFM to explore the interaction between spatial and channel information, enabling the learning of relevant pixel-to-category features. This aids in enhancing the model's ability to capture interclass and intraclass distinguishing features of objects in remote sensing images.

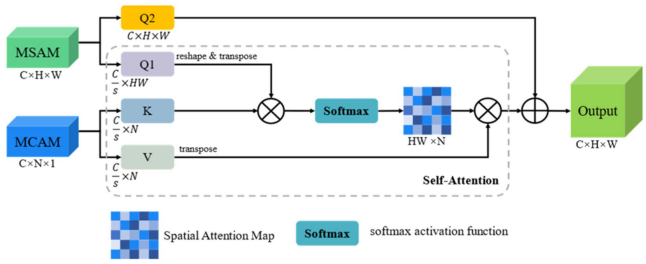


Fig. 5. Contextual information adaptive fusion module.

To deeply integrate contextual information from both channels and space, we use the spatial context information extracted by MSAM and the channel and category information extracted by MCAM as inputs to CIAFM. Inspired by the nonlocal self-attention mechanism in [49], we propose an improved lightweight self-attention mechanism. We use convolution with a kernel size of 3 to obtain $Q1 \in R^{\frac{C}{s} \times H \times W}$ and $Q2 \in R^{C \times H \times W}$ from the output of MSAM, where s is the reduction ratio of channels C. At the same time, we use convolution with a kernel size of 1 to obtain $K \in R^{\frac{C}{s} \times N}$ and $V \in R^{\frac{C}{s} \times N}$ from the output of MCAM. Then reshape $Q1$ to get $Q1 \in R^{\frac{C}{s} \times HW}$. The feature map $X_{qk} \in R^{HW \times N}$ containing the integrated channel and spatial context information is computed using (10). This process not only enables adaptive interaction between channel context information and spatial context information but also allows category information to guide pixel classification

$$X_{qk} = \sigma\left(\frac{Q^T K}{\sqrt{\frac{C}{s}}}\right) \tag{10}$$

where σ serves as the activation function, T represents transpose, C is the number of channels, and s is the reduction ratio of channel C. To further enhance the representation of spatial, channel, pixel, and category features, we multiply X_{qk} with V to obtain the feature details $X_{\text{attn}} = R^{\frac{C}{s} \times HW}$. Then, X_{attn} undergoes convolution with a kernel size of 1 and reshaping to obtain $X_{\text{attn}} \in R^{C \times H \times W}$. Finally, we concatenate X_{attn} with $Q2$ along the channel dimension, then we update the weights through convolution with a kernel size of 1 to obtain the output feature information $X_{\text{CIAFM}} \in R^{C \times H \times W}$ of CIAFM. In this module, we perform a linear transformation on the extracted channel and category information to use as the K and V in the self-attention mechanism. This operation reduces the computational cost of the self-attention module. Specifically, the computational cost of the SSAM is $C \times HW \times HW$ [49], where C represents the number of channels, H means the height of the input feature, and W denotes the width of the input feature. The computational cost of our proposed self-attention mechanism is $C \times N \times HW$, where N is the number of categories. Because the number of object categories in high-resolution remote sensing images is generally between 3 and 8, N is much smaller than HW. Therefore, the self-attention mechanism proposed in this article might significantly reduce the computational and parameter overhead of the network.

CIAFM not only adaptively integrates long-range information extracted from channels and spatial features but also learns the correlation between categories and pixels, leading to further improvement in the segmentation accuracy of remote sensing images.

IV. Dataset and Setting

A. Dataset

In this article, the following three datasets are used to validate the effectiveness of the proposed method.

ISPRS Vaihingen dataset: This dataset, provided by ISPRS, was collected from aerial imagery of the German city of Vaihingen and consists of 33 high-resolution images. Each image has an average size of about 2494 × 2064 pixels, contains red, near-infrared, and green bands, has a sampling distance of 9 cm on the ground, and contains six categories: impervious surface, building, low vegetation, tree, car, and background. We use images with ID: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38 as the test set; images with ID: 30 as the validation set and the rest of the images as the training set. The original images were cropped to 512×512 pixels size.

ISPRS Potsdam dataset: This dataset, provided by ISPRS, was collected from aerial imagery of the German city of Potsdam, with a total of 38 high-resolution images. The Potsdam dataset is available in the near-infrared, red, green, and blue bands, and the imagery is annotated with six categories: impervious surface, building, low vegetation, tree, car, and background. We use images with IDs: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13 as the test set, image with ID: 2_10 as the validation set, and the rest of the 22 images (except for the incorrectly annotated 7_10 image) as the training set. In our experiments, only three multispectral bands (red, green, and blue) were used. Each original image was cropped to 512 × 512 pixel size.

Multiobject coastal supervision semantic segmentation dataset (MO-CSSSD): This dataset is collected from aerial imagery of the coastal areas in southern China, used for coastal ecological environment monitoring. The dataset contains a total of 10 574 RGB images with a resolution of 512×512 pixels, including four categories: mangrove, aquaculture raft, aquaculture pond, and background. The spatial resolution of the RGB images is 0.58 m. According to the standard of dividing datasets in machine learning, this dataset is divided into a training set, validation set, and test set at a ratio of 6:2:2, that is 1100 images for the test set, 1100 images for the validation set, and 8734 images for the training set.

B. Experimental Settings

All experiments were implemented using the deep-learning framework PyTorch and run on NVIDIA GTX 3090 GPUs. To make the experiments converge quickly, we used AdamW as the training optimizer with an initial learning rate of $6e-5$, and adopted a cosine strategy to optimize the learning rate.

C. Evaluation Metrics

In the experiments, overall accuracy (OA), mean intersection over union (mIoU), and mean F1 score (mF1) were used as evaluation metrics to assess the performance of the proposed

TABLE I
ABLATION STUDY OF EACH COMPONENT OF THE AANET

Dataset	Method	OA	mF1	mIoU
Vaihingen	Baseline	90.08	87.12	77.71
	Baseline + MSAM	90.68	88.47	79.78
	Baseline + MSAM + MCAM	91.38	89.34	81.07
	Baseline + MSAM + MCAM+ CIAFM	**92.06**	**90.64**	**83.17**
Potsdam	Baseline	88.81	90.30	82.50
	Baseline + MSAM	89.25	90.96	83.64
	Baseline + MSAM + MCAM	89.88	91.39	84.33
	Baseline + MSAM + MCAM+ CIAFM	**90.84**	**92.16**	**85.67**

The best values in the column are in bold.

network. The equations for calculating each evaluation metric are as follows:

$$\mathrm{OA} = \frac{\sum_{k=1}^{N} \mathrm{TP}_k}{\sum_{k=1}^{N} (\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{TN}_k + \mathrm{FN}_k)} \tag{11}$$

$$\mathrm{mIoU} = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{12}$$

$$\mathrm{mF1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \times \frac{1}{N} \tag{13}$$

where TP_k, FP_k, TN_k, and FN_k indicate the true positive, false positive, true negative, and false negative, respectively, for objects indexed as class *k*. OA is calculated for all categories, including the background.

V. RESULTS AND DISCUSSION

A. Ablation Study

1) Components of AANet: To validate the performance of each component proposed in this article, we conducted a series of ablation experiments on the ISPRS Vaihingen and ISPRS Potsdam datasets. In this round of experiments, all parameters remained consistent except for the network structure.

Baseline: The baseline network is constructed using the ResNet backbone as part of the proposed network in this article. It solely extracts contextual information during the encoding process and progressively combines with skip connections during the decoding phase before upsampling.

MSAM: Four MSAMs were incorporated into the baseline. As shown in Table I, when MSAM is incorporated into the baseline network, there is a noticeable improvement in evaluation metrics, such as mIoU. The improvement in the ISPRS Vaihingen and ISPRS Potsdam datasets are 2.07% and 1.14%, respectively. These results demonstrate that MSAM, by capturing contextual information from multiple dimensions in the spatial domain and modeling the correlations between them, provides crucial context for the model, leading to improved segmentation accuracy.

The MCAM is added to the Baseline + MSAM, resulting in an average 1% improvement in mIoU. The MCAM eliminates redundant noise within channels, extracts contextual and class information, and enhances the network's feature representation capability. Significant improvements are also observed in mF1 and mIoU compared to the Baseline + MSAM ablation experiment.

CIAFM: We inserted a few CIAFMs into the Baseline + MSAM + MCAM to generate the entire AANet (indicated as the Baseline + MSAM + MCAM + CIAFM). As shown in Table I, adding the CIAFM module led to an increase of at least 2% in mIoU, demonstrating the effectiveness of the proposed CIAFM.

2) Attention Module Comparison: We conducted extensive attention module ablation experiments on the ISPRS Vaihingen dataset to verify the effectiveness of the attention module proposed in this article. We selected the same baseline with the previous section as the backbone.

To visually demonstrate the improvement brought by our proposed MCAM, we replaced the MCAM with the channel attention module from scSE [50] and the channel attention module proposed by ECANet, respectively. After replacement, we denoted the two networks as Baseline + scSE + MSAM + CIAFM and Baseline + ECA + MSAM + CIAFM, respectively. As shown in Table II, the experimental results indicate that our proposed network outperforms the Baseline + scSE + MSAM + CIAFM and the Baseline + ECA + MSAM + CIAFM in all three metrics. The channel attention presented by scSE extracts contextual information across the entire channel, which may lead to many irrelevant feature details within the channel, subsequently affecting the segmentation performance. MCAM utilizes 1-D convolution to capture local neighboring context and category features, eliminating redundant feature details within the channel. This confers an advantage in image segmentation tasks. The channel attention module proposed by ECA module extracts locally adjacent channel contextual information, which can avoid accumulating redundant feature details within the channel. However, by adopting a single-scale strategy to extract target features from images, it lacks the ability to adaptively perceive the significant variations in target features caused by the diverse scale differences in remote sensing images. Our proposed MCAM adopts a multiscale strategy to extract adjacent channel contextual information, addressing the limitations encountered by ECA module. Therefore, MCAM exhibits greater potential in perceiving significant variations in target features with diverse scale differences.

Similarly, we replaced the MSAM module proposed in this article with the spatial attention module introduced by coordinate

TABLE II
ABLATION STUDIES OF DIFFERENT ATTENTION MECHANISMS ON THE ISPRS VAIHINGEN DATASET

Attention method	OA	mF1	mIoU
Baseline+ scSE [50] + MSAM+ CIAFM	91.73	89.83	81.88
Baseline+ ECA [35] + MSAM+ CIAFM	91.73	89.91	82.02
Baseline+ MCAM+ Coordinate [51] + CIAFM	91.50	89.44	81.26
Baseline+ MCAM+ MSAM+ SUM	91.33	89.51	81.38
Baseline+ MCAM+ MSAM + SSAM [22]	91.76	90.02	82.19
Baseline+ CBAM [38]	91.87	90.16	82.39
Baseline+ GLTB [41]	91.37	89.33	81.13
Baseline+ MCAM+ MSAM+ CIAFM (ours)	**92.06**	**90.64**	**83.17**

The best values in the column are in bold.

[51], resulting in a comparison network, Baseline + MCAM + Coordinate + CIAFM. The experimental results, as shown in Table II, demonstrate that our proposed MSAM, when compared with the method after replacement, has improved by 0.56%, 1.20%, and 1.91% in terms of OA, mF1, and mIoU, respectively. The spatial attention module presented in coordinate calculates attention weights by introducing coordinate information, which facilitates the model in extracting crucial features from critical regions. However, this module presents challenges in getting further information from various spatial dimensions, leading to a lack of retention of spatial details. Additionally, the introduction of extensive coordinate information increases the computational and parameter load of the model. MSAM, through pooling and dot-product operations, extracts and adaptively fuses features from various dimensions in the spatial domain, thereby retaining richer spatial details. Consequently, in image semantic segmentation tasks, MSAM can effectively provide crucial position features for the model.

To validate the effectiveness of the proposed CIAFM self-attention mechanism, we replaced CIAFM with a feature map summation operation and constructed the Baseline + MCAM + MSAM + SUM network. The experimental results, as shown in Table II, indicate that compared to AANet, the Baseline + MCAM + MSAM + SUM network shows a decrease of 1.79% in mIoU, demonstrating that the summation operation cannot fully leverage the spatial and channel contextual information, thereby limiting the segmentation performance. In the same way, we replaced CIAFM with a SSAM [22] and constructed a new network: Baseline + MCAM + MSAM + SSAM. The experimental results, as shown in Table II, indicate that compared to AANet, the Baseline + MCAM + MSAM + SSAM network has room for improvement in all evaluation metrics. This is because CIAFM utilizes category information to guide pixel-level classification within the spatial context. Above all, the results demonstrate that our proposed CIAFM not only fully leverages the correlation between space and channels, but also indicates that CIAFM can utilize category information existing between channels to guide spatial pixel classification.

The network proposed in this article adopts a hybrid architecture, which includes MCAM, MSAM, and CIAFM. To validate whether this modular design approach can effectively improve the image segmentation performance, we replaced the MCAM, MSAM, and CIAFM with the hybrid channel-attention and spatial attention module proposed by CBAM and the global–local transformer block proposed by UNetFormer, respectively. As a result, we obtained two new networks: Baseline + CBAM [38] and Baseline + GLTB [41]. The experimental results in Table II demonstrate that, while the Baseline + CBAM network has achieved satisfactory results, the sequential connection of channel and spatial attention in CBAM leads to the neglect of spatial details when extracting channel context information and the loss of channel features when extracting spatial context information. This might cause inadequate output to leverage the relevant relationship between space and channels. Therefore, compared to AANet, the capability of Baseline + CBAM to extract contextual information needs to be improved. Compared to Baseline + GLTB, AANet showed improvements of 0.69%, 1.31%, and 2.04% in OA, mF1, and mIoU, respectively. In conclusion, this article has demonstrated that the attention modules in AANet can fully leverage the contextual information of both space and channels and effectively extract and fuse the correlated relationships between space and channels.

3) Encoder Module Comparison: To explore the impact of encoders on the overall network, we replaced the ResNet18 encoder of the proposed network with vision transformer and ResNet50, respectively, constructing two networks: VITAANet and ReAANet. The experimental results are shown in Table III. As shown in Table III, the networks using vision transformer and ResNet50 as encoders achieved better performance, albeit with increased network complexity. For example, on the ISPRS Vaihingen dataset, VITAANet achieved mIoU, mF1, and OA scores that were 0.77%, 0.47%, and 0.23% higher than AANet, respectively. However, the number of parameters and FLOPs were also higher by 94.21M and 160.93G, respectively. Balancing network accuracy and complexity, we ultimately chose ResNet18 as the network's encoder.

B. Comparison of Computing Complexity

In practical applications, algorithm complexity is an essential metric for evaluating its performance. We compared AANet with lightweight networks based on parameter count, computational complexity, mF1, and mIoU on the ISPRS Vaihingen test set. The comparison results are shown in Table IV. Compared to the lightweight network BiSeNetV2 [52], which has the most minor parameter count and computational complexity, our network shows significant improvements in OA and mIoU. Specifically, on the ISPRS Vaihingen dataset, we achieved an increase of

TABLE III
ABLATION STUDY ON DIFFERENT ENCODERS ON THE ISPRS VAIHINGEN TEST SET AND ISPRS POTSDAM TEST SETS

Method	Backbone	Parameters(M)	FLOPs(G)	OA	mF1	mIoU
VITAANet	ViT	106.93	176.80	92.29/**91.56**	**91.09/92.93**	**83.94/87.01**
ReAANet	ResNet50	26.71	31.32	**92.35**/91.50	91.08/92.83	83.90/86.84
AANet	ResNet18	**12.72**	**15.87**	92.06/90.84	90.64/92.16	83.17/85.67

The best values in the column are in bold.

TABLE IV
QUANTITATIVE COMPARISON RESULTS ON THE ISPRS VAIHINGEN TEST SET AND THE ISPRS POTSDAM TEST SET WITH STATE-OF-THE-ART LIGHTWEIGHT NETWORKS

Method	Backbone	Parameters(M)	FLOPs(G)	OA	mF1	mIoU
DANet [20]	ResNet18	12.59	74.23	90.15/89.95	89.04/91.36	80.60/84.30
PSPNet [14]	ResNet18	24.42	98.21	91.28/89.60	89.23/91.01	80.91/83.71
BiSeNetV1 [53]	ResNet18	40.90	23.10	91.07/89.69	88.32/91.00	79.53/83.70
BiSeNetV2 [52]	–	**5.10**	**11.19**	90.18/88.15	86.87/88.93	77.35/80.25
MANet [37]	ResNet18	11.98	22.19	91.43/90.35	89.87/91.70	81.93/84.88
ABCNet [42]	ResNet18	13.39	15.62	91.78/90.33	89.82/91.51	81.84/84.58
BotNet [54]	ResNet18	13.05	11.44	91.05/89.72	87.91/91.03	78.96/83.79
UNetFormer [41]	ResNet18	11.68	11.74	91.81/90.75	90.34/92.07	82.69/85.51
AANet (ours)	ResNet18	12.72	15.87	**92.06/90.84**	**90.64/92.16**	**83.17/85.67**

The best values in the column are in bold.

1.88% in OA, 3.77% in mF1, and 5.82% in mIoU. On the ISPRS Potsdam dataset, we gained a rise of 1.69% in OA, 3.23% in mF1, and 5.42% in mIoU. Furthermore, the proposed method in this article outperforms the state-of-the-art lightweight network UNetFormer on two public datasets. The AANet network utilizes a multiscale strategy to extract contextual information from neighboring channels, eliminating redundant noise within the channels. It employs pooling and dot-product operations to extract and adaptively fuse spatial details across dimensions, providing crucial feature information for the model. The improved self-attention mechanism is used to explore the complementary nature of channel and spatial contextual information, enhancing the feature representation capability of the network. Therefore, compared to networks of similar scale, the proposed AANet demonstrates significant improvements in segmentation accuracy.

C. Quantitative Comparison of Diverse Lightweight Methods

We compared our proposed method with eight excellent lightweight networks on the ISPRS Vaihingen and ISPRS Potsdam test sets. These networks include: DANet [20], which uses a parallel dual attention mechanism to extract spatial and channel context information; PSPNet [14], which employs pyramid pooling modules to capture feature information at different scales in images; BiSeNet_V1 [53], featuring a dual-stream network structure with spatial and context paths; BiSeNet_V2 [52], which introduces more cross-modal information and effective feature fusion mechanisms in the dual-stream network architecture; MANet [37], which develops a novel kernel attention mechanism for extracting context information in images; ABCNet [42], which preserves spatial detail information and image context information separately using a dual-path approach; BotNet [54], which replaces spatial convolutions with global attention in the last three bottleneck blocks of ResNet; and UNetFormer [41], which designs a brand new transformer decoder and utilizes feature refinement head (FRH) to refine output features.

The experimental results on the ISPRS Vaihingen dataset are shown in Table V. Due to the utilization of class information to guide pixel classification within the CIAFM and the deep integration of contextual information across channels and spatial dimensions, we have effectively leveraged the interrelationships between different contextual features. As a result, our method has achieved optimal performance in distinguishing between the challenging categories of trees and cars, yielding respective accuracies of 81.53% and 79.39%.

Because of adopting a dual-branch approach without interaction in BiSeNet_V2, which extracts detail and semantic information separately before fusing them, the effectiveness of this method in extracting image information is somewhat unsatisfactory. In this article, the decoder part utilizes the CIAFM at each stage to deeply integrate and extract the complementarity between spatial and channel contexts, making full use of the hierarchically extracted features from the encoder. Thus, the method proposed in this article scores high compared to BiSeNet_V2, the lightest of the comparative networks in terms of IoU on each category, and overall metrics. Compared to UNetFormer, the lightweight network that achieved the best performance last year, our approach shows an improvement of 0.25% in OA, 0.30% in mF1, and 0.48% in mIoU. Additionally, compared to ABCNet, MANet, and BotNet of the same scale, our approach exhibits an average improvement of 1.2% in IoU values for each category and significant enhancements in terms of mF1, OA, and mIoU.

In addition, we provide visual experimental results on the ISPRS Vaihingen dataset, as shown in Fig. 6. From the images in the first, second, and third rows in Fig. 6, it can be observed that our proposed method performs remarkably well in segmenting

TABLE V
QUANTITATIVE COMPARISON RESULTS ON THE ISPRS VAIHINGEN TEST SET WITH STATE-OF-THE-ART LIGHTWEIGHT NETWORKS

Method	Backbone	Impervious surfaces	Building	Low vegetation	Tree	Car	OA	mF1	mIoU
DANet [20]	ResNet18	90.01	89.23	70.90	80.32	72.54	91.15	89.04	80.60
PSPNet [14]	ResNet18	90.17	89.21	71.02	80.89	73.26	91.28	89.23	80.91
BiSeNetV1 [53]	ResNet18	90.06	88.67	70.81	80.11	68.01	91.07	88.32	79.53
BiSeNetV2 [52]	–	88.61	86.39	69.56	79.48	62.70	90.18	86.87	77.35
MANet [37]	ResNet18	90.56	90.23	70.56	80.68	77.59	91.43	89.87	81.93
ABCNet [42]	ResNet18	90.88	89.72	**72.82**	81.49	74.30	91.78	89.82	81.84
BotNet [54]	ResNet18	89.98	88.83	70.71	79.98	65.30	91.05	87.91	78.96
UNetFormer [41]	ResNet18	91.05	90.79	71.61	81.16	78.85	91.81	90.34	82.69
AANet (ours)	ResNet18	**91.50**	**91.09**	72.35	**81.53**	**79.39**	**92.06**	**90.64**	**83.17**

The best values in the column are in bold.

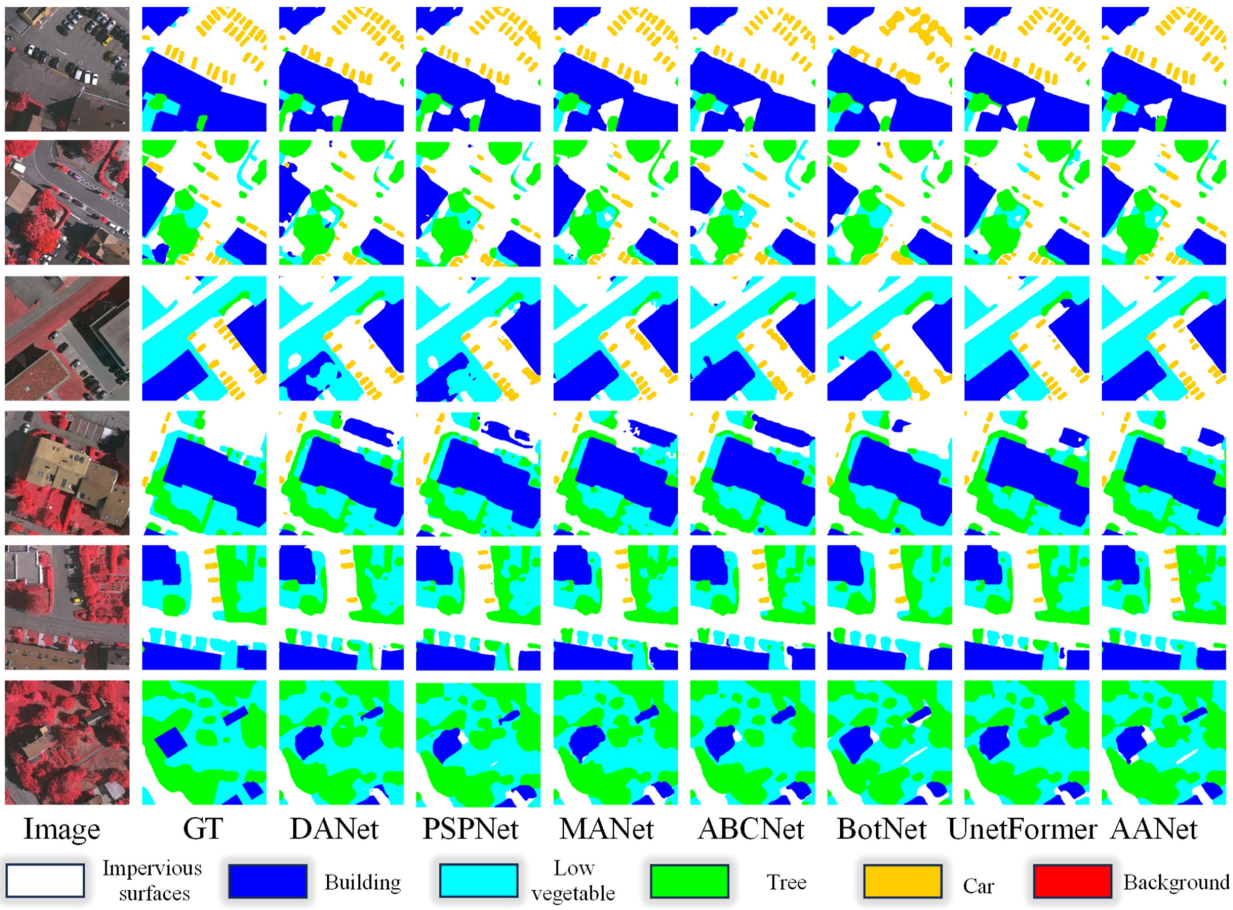


Fig. 6. Results of visualization experiments on the ISPRS Vaihingen dataset.

the highly challenging car category. In the examples in the last three rows, it is evident that our proposed method more easily distinguishes between tree and low vegetation categories. This is because our strategy utilizes the category information contained in the channels to guide the classification of pixels in space, thus better capturing the objects' shape, contour, and other features during segmentation. Our method utilizes a multiscale strategy to extract contextual information from neighboring channels, effectively eliminating redundant channel noise.

Table VI presents the experimental results on the ISPRS Potsdam dataset. It is evident that our method has achieved significant results across various evaluation metrics. Specifically, compared to the computationally and parameterwise minimal BiSeNet_V2, our network achieved the highest scores for each category, demonstrating that our network has improved segmentation performance with an appropriate increase in network complexity. In comparison to networks with roughly the same complexity, such as ABCNet, MANet, DANet, and BotNet, our

TABLE VI
QUANTITATIVE COMPARISON RESULTS ON THE ISPRS POTSDAM TEST SET WITH STATE-OF-THE-ART LIGHTWEIGHT NETWORKS

Method	Backbone	Impervious surfaces	Building	Low vegetation	Tree	Car	OA	mF1	mIoU
DANet [20]	ResNet18	86.18	90.43	75.83	78.30	90.78	89.95	91.36	84.30
PSPNet [14]	ResNet18	85.84	89.38	74.86	78.24	90.25	89.60	91.01	83.71
BiSeNetV1 [53]	ResNet18	85.97	90.65	75.84	77.02	89.05	89.69	91.00	83.70
BiSeNetV2 [52]	-	83.83	88.14	72.70	74.78	81.79	88.15	88.93	80.25
MANet [37]	ResNet18	86.15	91.50	76.69	79.03	91.04	90.35	91.70	84.88
ABCNet [42]	ResNet18	86.74	91.62	76.15	78.04	90.33	90.33	91.51	84.58
BotNet [54]	ResNet18	86.49	91.37	75.15	76.49	89.45	89.72	91.03	83.79
UNetFormer [41]	ResNet18	**87.69**	**92.05**	77.04	79.32	91.47	90.75	92.07	85.51
AANet (ours)	ResNet18	87.67	91.93	**77.19**	**79.73**	**91.80**	**90.84**	**92.16**	**85.67**

The best values in the column are in bold.

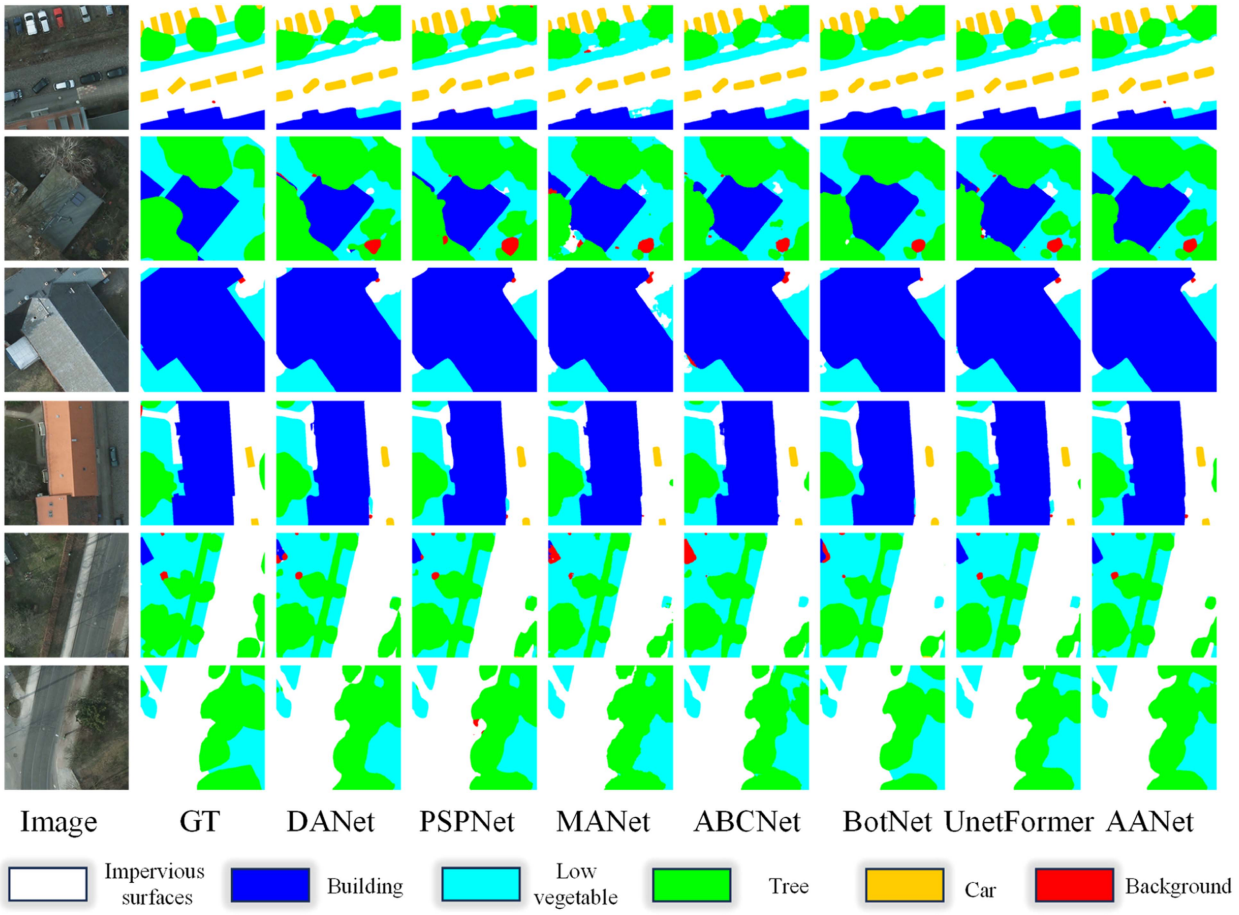


Fig. 7. Results of visualization experiments on the ISPRS Potsdam dataset.

proposed network demonstrates superiority. While UNetFormer, which achieved the best performance among lightweight networks last year, exhibits good segmentation performance, our network surpasses it in most evaluation metrics. Due to the feature refinement module in UNetFormer, which integrates semantic and spatial detail information at a deep level, the network reduces the semantic gap between the two features and further improves semantic accuracy. As a result, the network we propose still has room for improvement.

We have provided visual experimental results, as shown in Fig. 7, from which it is evident that the segmentation performance of the proposed method in this article is superior to other networks. As depicted in the example images from the first to the fourth rows in Fig. 7, our proposed method excels in extracting comprehensive feature information for cars, low vegetation, and trees, leading to a more precise segmentation. This can be attributed to our adoption of a multiscale strategy, which enables adaptive extraction of feature information from

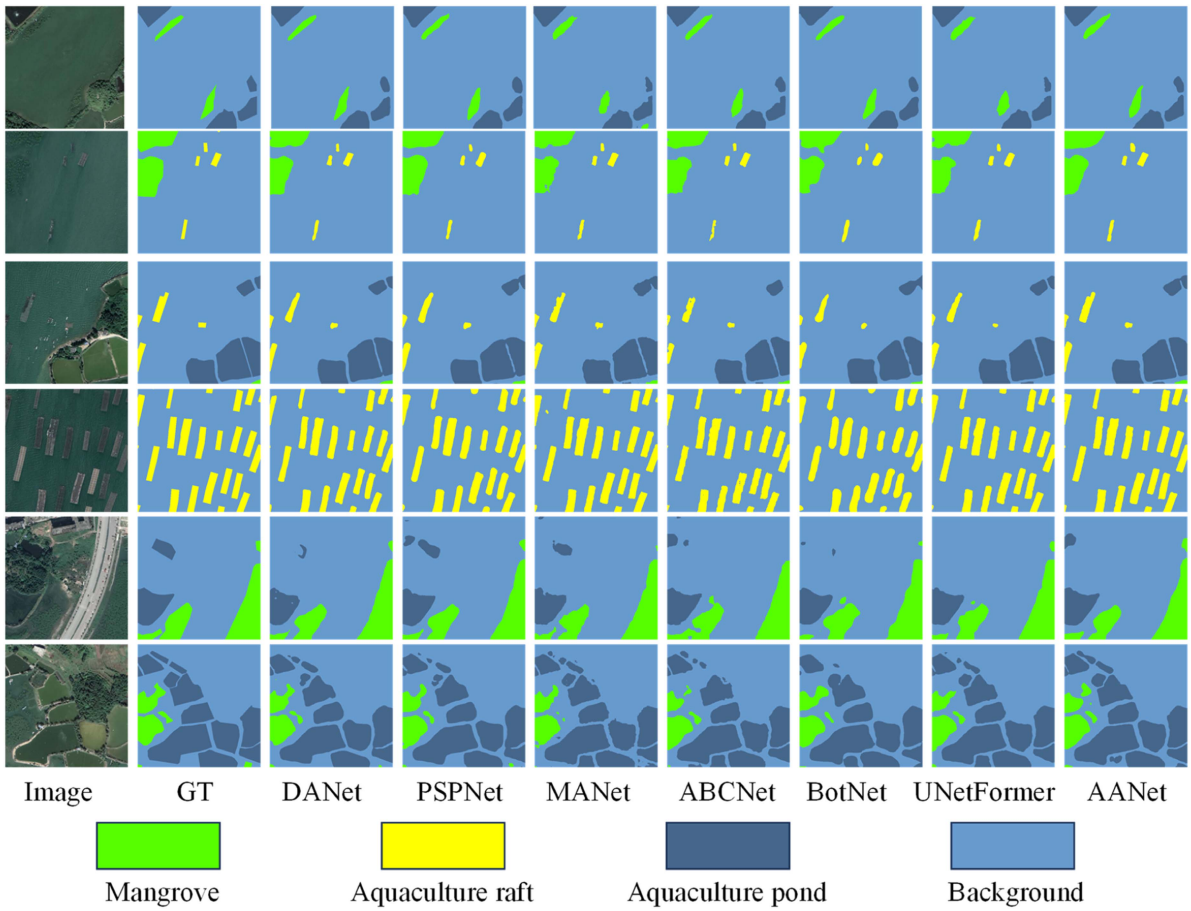


Fig. 8. Results of visualization experiments on the MO-CSSSD dataset.

TABLE VII
QUANTITATIVE COMPARISON RESULTS ON THE MO-CSSSD TEST SET WITH STATE-OF-THE-ART LIGHTWEIGHT NETWORKS

Method	Backbone	Mangrove	Aquaculture raft	Aquaculture pond	Background	OA	mF1	mIoU
DANet [20]	ResNet18	89.59	86.86	80.16	98.50	98.61	93.93	88.78
PSPNet [14]	ResNet18	88.95	86.77	81.00	98.50	98.61	93.95	88.81
BiSeNetV1 [53]	ResNet18	88.55	86.09	76.55	98.26	98.38	93.07	87.36
BiSeNetV2 [52]	-	82.37	47.46	63.53	96.90	97.09	82.71	72.56
MANet [37]	ResNet18	87.73	89.70	79.38	98.36	98.48	93.93	88.79
ABCNet [42]	ResNet18	87.94	87.51	76.61	98.25	98.37	93.20	87.58
BotNet [54]	ResNet18	84.83	82.60	75.76	98.09	98.22	91.88	85.32
UNetFormer [41]	ResNet18	88.53	**89.91**	79.00	98.42	98.53	94.02	88.96
AANet (ours)	ResNet18	**88.97**	89.82	**81.38**	**98.53**	**98.63**	**94.45**	**89.68**

The best values in the column are in bold.

remote sensing images. In less complex settings, our method demonstrates enhanced capability in distinguishing between the categories of low vegetation and trees, as illustrated in the last two rows of Fig. 7. This is facilitated by our proposed network's effective utilization of the correlation between channels and spatial information, thereby effectively mitigating interference from background features.

Table VII presents the experimental comparison results of each model on the MO-CSSSD test set. As shown in Table VII, the AANet method proposed in this article achieves the best experimental results in most aspects. For example, it obtains mIoU scores of 88.97%, 81.38%, and 98.53% for the mangrove, aquaculture pond, and background categories, respectively. Compared with ABCNet, MANet, BotNet, and DANet

TABLE VIII
QUANTITATIVE COMPARISON RESULTS ON THE ISPRS VAIHINGEN TEST SET AND THE ISPRS POTSDAM TEST SET WITH STATE-OF-THE-ART LIGHTWEIGHT NETWORKS

Method	Backbone	Impervious surfaces	Building	Low vegetation	Tree	Car	OA	mF1	mIoU
UNetFormer	ResNet18	**93.56/87.94**	90.23/**92.78**	**72.55/77.55**	80.72/79.48	76.85/**92.73**	**92.82/91.04**	90.37/**92.40**	82.78/**86.10**
AANet	ResNet18	93.38/86.54	**90.32**/92.21	70.93/77.53	**81.02/79.75**	**79.29**/92.33	92.74/90.62	**90.49**/92.16	**82.99**/85.67

The experimental results were measured using a 1024 × 1024 input.

methods of the same volume, our AANet achieved the highest score. As shown in Table VII, the UNetFormer, which developed a novel transformer encoder, achieved an mIoU of 89.91% in the aquaculture raft category. This is because UNetFormer optimized the feature information of the final output by utilizing the FRH. It is noteworthy that our proposed AANet falls short of the best value by 0.09% in the aquaculture raft category. However, it achieved the best results in the other three categories and in overall OA, mF1, and mIoU.

The visualization results of DANet, PSPNet, MANet, ABCNet, BotNet, UNetFormer, and AANet on the MO-CSSSD test set are shown in Fig. 8. From the legends in the first, second, and fifth rows of Fig. 8, it is evident that the AANet proposed in this article produces segmentation contours in the aquaculture pond category that are closer to the real labels. This benefit comes from our proposed MCAM module, which adopts convolution with kernel sizes adaptive to channels to extract multiscale information from images. From the legends in the third, fourth, and fifth rows of Fig. 8, it is clearly visible that the AANet method proposed in this article achieves better segmentation effects in the mangrove and aquaculture raft categories, with clearer contours. These improvements in segmentation results are primarily due to the category information extracted from channels guiding the pixel information extracted in space. The superiority of our method is demonstrated through the various legends in Fig. 8.

We conducted comparative experiments by providing larger sized images used in UNetFormer as network inputs, and the results are shown in Table VIII. UNetFormer proposes the FRH that narrows the semantic gap between the rich spatial information features output from the first stage of the encoder and the deep global and local semantic information features. This results in superior performance in segmenting certain categories compared to our method. However, our method can adaptively capture category information from channels, which makes it perform exceptionally well on tree and car categories. While larger inputs will likely help improve accuracy in UNetFormer, they may also result in higher computational effort.

VI. CONCLUSION

This article proposes the AANet for semantic segmentation in high-resolution remote sensing images. Due to the fact that contextual information enhances the ability to identify object categories in semantic segmentation tasks, we designed the MCAM. This module utilizes a multiscale strategy to extract contextual information from adjacent channels, enabling the perception of features from objects of different sizes and eliminating redundant noise within the channels. By incorporating MCAM, our network can better capture the characteristics of objects at different scales and effectively suppress the influence of noise. In addition, we also developed the MSAM. MSAM utilizes pooling and dot product operations to extract and fuse contextual information from diverse dimensions in the spatial domain, which increases the model's attention to critical regions and suppresses interference from the background. To effectively utilize the correlation between contextual information, we proposed the CIAFM. This module employs an improved self-attention mechanism to deeply integrate channel and spatial contextual information while reducing network complexity. This enables the network to be used for real-time segmentation of high-resolution images. In future article, we will continue to explore the contextual information and its correlations present in high-resolution images.

REFERENCES

[1] R. Li, S. Y. Zheng, C. X. Duan, L. B. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-Spatial Inf. Sci.*, vol. 25, no. 2, pp. 278–294, Apr. 2022, doi: 10.1080/10095020.2021.2017237.

[2] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017, doi: 10.1109/TGRS.2016.2612821.

[3] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018, doi: 10.1016/j.isprsjprs.2018.01.021.

[4] D. Griffiths and J. Boehm, "Improving public data for building segmentation from convolutional neural networks (CNNs) for fused airborne lidar and image data using active contours," *ISPRS J. Photogrammetry Remote Sens.*, vol. 154, pp. 70–83, 2019, doi: 10.1016/j.isprsjprs.2019.05.013.

[5] P. Shamsolmoali, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, "Road segmentation for remote sensing images using adversarial spatial pyramid networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4673–4688, Jun. 2021, doi: 10.1109/TGRS.2020.3016086.

[6] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1873–1876, doi: 10.1109/IGARSS.2015.7326158.

[7] A. Samie et al., "Examining the impacts of future land use/land cover changes on climate in Punjab province, Pakistan: Implications for environmental sustainability and economic growth," *Environ. Sci. Pollut. Res.*, vol. 27, no. 20, pp. 25415–25433, Jul. 2020, doi: 10.1007/s11356-020-08984-x.

[8] H. Yin, D. Pflugmacher, A. Li, Z. Li, and P. Hostert, "Land use and land cover change in Inner Mongolia—Understanding the effects of China's re-vegetation programs," *Remote Sens. Environ.*, vol. 204, pp. 918–930, 2018, doi: 10.1016/j.rse.2017.08.030.

[9] C. Zhang, P. A. Harrison, X. Pan, H. Li, I. Sargent, and P. M. Atkinson, "Scale sequence joint deep learning (SS-JDL) for land use and land cover classification," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111593, doi: 10.1016/j.rse.2019.111593.

[10] C. Zhang et al., "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, 2019, doi: 10.1016/j.rse.2018.11.014.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[13] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[16] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 783–792.

[17] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974, doi: 10.1109/T-C.1974.223784.

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 603–612.

[19] J. Jiang, X. Feng, Q. Ye, Z. Hu, Z. Gu, and H. Huang, "Semantic segmentation of remote sensing images combined with attention mechanism and feature enhancement U-Net," *Int. J. Remote Sens.*, vol. 44, no. 19, pp. 6219–6232, Oct. 2023, doi: 10.1080/01431161.2023.2264502.

[20] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[21] X. Hu, P. Zhang, Q. Zhang, and F. Yuan, "GLSANet: Global-local self-attention network for remote sensing image semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6000105, doi: 10.1109/LGRS.2023.3235117.

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[25] J. Nie, L. Huang, C. Zheng, X. Lv, and R. Wang, "Cross-scale graph interaction network for semantic segmentation of remote sensing images," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 6, pp. 1–8, 2023, doi: 10.1145/3558770.

[26] K.-H. Liu and B.-Y. Lin, "MSCSA-Net: Multi-scale channel spatial attention network for semantic segmentation of remote sensing images," *Appl. Sci.*, vol. 13, no. 17, 2023, Art. no. 9491.

[27] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020, doi: 10.1016/j.isprsjprs.2020.01.013.

[28] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[29] Z. Xu, W. Zhang, T. Zhang, and J. Li, "HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 13, no. 1, Dec. 2020, Art. no. 71, doi: 10.3390/rs13010071.

[30] X. Qi, Y. Mao, Y. Zhang, Y. Deng, H. Wei, and L. Wang, "PICS: Paradigms integration and contrastive selection for semisupervised remote sensing images semantic segmentation," *IEEE Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602119, doi: 10.1109/tgrs.2023.3239042.

[31] X. Zhang, Z. Wang, J. Zhang, and A. Wei, "MSANet: An improved semantic segmentation method using multi-scale attention for remote sensing images," *Remote Sens. Lett.*, vol. 13, no. 12, pp. 1249–1259, Dec. 2022, doi: 10.1080/2150704X.2022.2142075.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[33] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605116, doi: 10.1109/TGRS.2023.3256064.

[34] J. Guo et al., "Spanet: Spatial pyramid attention network for enhanced image recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6, doi: 10.1109/ICME46284.2020.9102906.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.

[36] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8009205, doi: 10.1109/LGRS.2021.3063381.

[37] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713, doi: 10.1109/TGRS.2021.3093977.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[39] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603018, doi: 10.1109/tgrs.2021.3065112.

[40] H. Bi et al., "Not just learning from others but relying on yourself: A new perspective on few-shot segmentation in remote sensing," *IEEE Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5623621, doi: 10.1109/TGRS.2023.3326292.

[41] L. B. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022, doi: 10.1016/j.isprsjprs.2022.06.008.

[42] R. Li, S. Y. Zheng, C. Zhang, C. X. Duan, L. B. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021, doi: 10.1016/j.isprsjprs.2021.09.005.

[43] J. Long, M. Li, and X. Wang, "Integrating spatial details with long-range contexts for semantic segmentation of very high-resolution remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2501605, doi: 10.1109/LGRS.2023.3262586.

[44] J. Wang, X. L. Zhang, T. H. Yan, and A. H. Tan, "DPNet: Dual-pyramid semantic segmentation network based on improved deeplabv3 plus," *Electronics*, vol. 12, no. 14, Jul. 2023, Art. no. 3161, doi: 10.3390/electronics12143161.

[45] Q. Zhao, J. H. Liu, Y. W. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5403913, doi: 10.1109/tgrs.2021.3085889.

[46] B. K. Lin, G. Yang, Q. Zhang, and G. X. Zhang, "Semantic segmentation network using local relationship upsampling for remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8006105, doi: 10.1109/lgrs.2020.3047443.

[47] Z. M. Yu et al., "RSLC-Deeplab: A ground object classification method for high-resolution remote sensing images," *Electronics*, vol. 12, no. 17, Sep. 2023, Art. no. 3653, doi: 10.3390/electronics12173653.

[48] X. Xiong, X. P. Wang, J. H. Zhang, B. X. Huang, and R. F. Du, "TCUNet: A lightweight dual-branch parallel network for sea-land segmentation in remote sensing images," *Remote Sens.*, vol. 15, no. 18, Sep. 2023, Art. no. 4413, doi: 10.3390/rs15184413.

[49] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 593–602.

[50] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2018, pp. 421–429.

[51] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

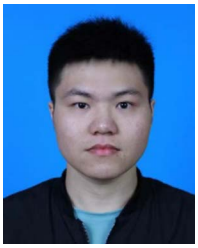
[52] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021, doi: 10.1007/s11263-021-01515-2.

[53] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

[54] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529.

Yan Chen received the M.Sc. degree in cartography and geographical information system from the China University of Mining and Technology, Xuzhou, China, in 2014, and the Ph.D. degree in spatial information management and modeling TU Dortmund, Dortmund, Germany, in 2019.

He studied spatial information management and modeling as a member of the Spatial Information Management and the Modelling Department of the School of Spatial Planning. He joined the Collaborative Innovation Centre for Computer Vision and Pattern Recognition, School of Artificial Intelligence and Big Data, Hefei University, Hefei, China, as Lecturer, in 2022. His research interests include remote sensing image analysis, computer vision and pattern recognition, as well as deep learning and optimization algorithms.



Qianchuan Zhang the B.Sc. degree in computer science from the Sichuan Minzu College, Kangding, China, in 2020. He is currently working toward the M.Sc. degree with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China.

His research interests include intelligent parsing of remote sensing images and deep learning.

Xiaofeng Wang received the the B.Sc. degree in software engineering from the Anhui University, Hefei, China, in 1999, the M.Sc. degree in pattern recognition and intelligent system from the University of Chinese Academy of Sciences, Beijing, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, in 2009.

He is currently a Professor with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, and is a provincial academic and technical leader reserve candidate. His research interests include computer vision and pattern recognition, and image processing.

Quan Dong received the B.Sc. degree in computer science and technology from Nanjing Normal University Zhongbei College, Zhenjiang, China, in 2022. He is currently working toward the M.Sc. degree in electronic information with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China.

His research interests include semantic segmentation of remote sensing images.

Menglei Kang received the B.Sc. degree in computer and information engineering from Chuzhou University, Chuzhou, China, in 2021, and the M.Sc. degree in electronic information from Hefei University, Hefei, China, in 2024.

His research interests include semantic segmentation of remote sensing images.

Wenxiang Jiang received the B.Sc. degree in computer and information engineering from the Anhui University of Technology, Ma'anshan, China, in 2021, and the M.Sc. degree in electronic information from Hefei University, Hefei, China, in 2024.

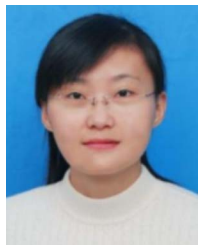
His research interests include computer vision in deep learning and remote sensing image analysis.

Mengyuan Wang received the B.Sc. degree in Internet of Things engineering from Applied Technology College of Soochow University, Jiangsu, China, in 2020 , and the M.Sc. degree in electronic information from Hefei University, Hefei, China, in 2024.

Her research interests include semantic segmentation of high-resolution remote sensing images and deep learning.

Lixiang Xu received the B.Sc. degree from Fuyang Normal University, Fuyang, China, in 2005, and M.Sc. degree from the Harbin University of Science and Technology, Harbin, China, in 2008, both in applied mathematics, and the Ph.D. degree in computer application technology from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2017.

He worked with the Huawei Technologies Co., Ltd., Shenzhen, China, in 2008, before joining the Hefei University, Hefei in the following year. He has been awarded a scholarship to pursue his study in Germany as a joint Ph.D. student from 2015 to 2017. He is currently a Postdoctoral Researcher with the University of Science and Technology of China, Hefei. His research interests include structural pattern recognition, machine learning, graph spectral analysis, and image and graph matching, especially in kernel methods and complexity analysis on graphs and networks.



Chen Zhang was born in Anhui, China. She received the M.S. degree in computational mathematics from the Anhui University, Hefei, China, in 2011, and the Ph.D. degree in information management and systems from the Hefei University of Technology, Hefei, in 2016.

She is currently an Associate Professor with the School of Artificial Intelligence and Big Data, Hefei University, Hefei. Her research interests include machine learning and artificial intelligence.