IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 August 2024, accepted 26 August 2024, date of publication 28 August 2024, date of current version 12 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3451153

APPLIED RESEARCH

MFAFNet: A Multiscale Fully Attention Fusion Network for Remote Sensing Image Semantic Segmentation

YUANYUAN DANG, YU GAO, AND BING LIU
School of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China
Corresponding author: Bing Liu (liubing@ccut.edu.cn)

ABSTRACT The semantic segmentation of high-resolution remote sensing images is widely used in various precision agriculture, urban planning, and environmental detection are some examples of these industries. Convolutional neural networks (CNNs) are excellent in the semantic segmentation of remote sensing images. CNN excels in extracting local feature details but lacks the ability to model global context data. Therefore, to obtain rich local-global information about context, we describe in this work a semantic segmentation network design technique for remote sensing, based on an encoder-decoder structure, which is named Multiscale Fully Attention Fusion Network for Remote Sensing Image Semantic Segmentation (MFAFNet). In particular, to improve the segmentation efficiency, the encoder's extractor of features was ResNet18, after which the explicit visual center module EVC and the full attention network FANB are intended to retrieve the detailed global context data. Finally, the gated channel attention fusion module (GCF) tries to augment channel interaction information in the decoder stage while fusing low-level characteristics for efficient aggregation. During our research and testing, we used the publicly available Vaihingen and Potsdam datasets from the International Society for Photogrammetry and Remote Sensing (ISPRS), as well as the LoveDA dataset. Meanwhile, it demonstrates that MFAFNet outperforms other well-liked methods in terms of competition. We further validated the efficiency of the network components in the study by conducting ablation experiments on the Vaihingen dataset.

INDEX TERMS Semantic segmentation, remote sensing, global-local context, attention mechanism.

I. INTRODUCTION

Given how quickly remote sensing technology is developing, obtaining images is becoming easier and easier. high-resolution remote sensing images, which provide a wealth of geographical and semantic information. Due to the quick growth of high-resolution remote sensing images, semantic segmentation has emerged as a key approach for feature recognition and area statistics of high-resolution remote sensing images [1]. Currently, urban planning [2], [3], precision agriculture [4], [5], [6], disaster assessment [7], [8], land resource management [9], [10], and environmental detection [11], [12], [13] depend heavily on the semantic segmentation of remote sensing images. Nevertheless, high-resolution remote sensing images include complex backgrounds. The majority of little targets, and also suffer from sample imbalance [14]. Traditional approaches to semantic segmentation, include thresholding [15], clustering-based methods [16], edge detection methods [17], and conditional random fields [18], [19], which mainly use shallow semantic information, such as color, and texture, for segmentation, have problems such as poor generalization ability [20] and insufficient deep feature extraction [21]. Exploration that relies solely on superficial information can no longer meet the requirements of remote sensing images in application areas because of the complexity of the background of remote sensing images. Therefore, raising the accuracy and efficiency of semantic segmentation in high-resolution remote sensing

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

images is indeed a challenging job. As deep learning develops greater popularity, specialists working on studies have developed various typical semantic segmentation models, which create a strong basis for remote sensing-based high-resolution semantic segmentation. Among them, Long et al. [22] suggested applying a fully convolutional network. As the first network model in the field of semantic segmentation, FCN uses an encoder-decoder structure to build a deep learning network. Additionally, it's the first attempt at semantic segmentation using CNN. It solved the pixel-level segmentation of images problem by converting the fully connected layer of CNN to a layer that uses convolution. After FCN, CNN-based methods dominate semantic segmentation of remotely sensed images [23], [24], [25], [26], but the FCN decoder is too simplified, which leads to low segmentation accuracy and easy loss of detailed information. Taking this into consideration, U-Net [27] proposed a working encoder-decoder arrangement. By fusing low and high-level characteristics using jump connections, we can increase the segmentation accuracy. In order to get back the structure of the spatial and contextual information lost in the feature map, the encoder-decoder is a derived model that uses an encoder to extract the details of the features of an image from the feature map and a decoder to reassemble this feature information. After that, the encoder-decoder framework becomes the remote sensing image segmentation mainstream structure [28], [29]. In an effort to increase the segmentation accuracy and inference speed of high-resolution remote sensing images, Wang et al. [30] created a lightweight encoder-decoder structure. Since remotely sensed images are characterized by complex backgrounds, small targets, and fuzzy boundaries, there have been significant advancements in the field of semantic segmentation of remotely sensed images using convolutional neural networks (CNNs) as the codec approach. Cui et al. [31] proposed a U-Net based semantic segmentation method that incorporates a channel attention mechanism and sub-pixel convolution method in the encoder-decoder to better capture feature information. However, the convolutional operation of CNNs with fixed sense fields leads to a lack of ability to extract global contextual information and to model distant dependencies, and the segmentation results are often ambiguous if only local information is modeled in semantic segmentation tasks [32].

To address the problems mentioned above, ERFNet [33] uses residual concatenation and decomposition convolution to ensure segmentation accuracy while improving inference speed. BiSeNet [34] proposes a bilateral segmentation network that combines spatial information and contextual paths. DABNet [35] extracts local information and contextual information simultaneously with asymmetric convolution and dilation convolution with comparable accuracy and inference speed. SPANet [36] solved the foreground-background imbalance problem by designing spatial adaptive convolution in the decoder. These networks though achieve a certain balance between performance and speed. But still, they cannot get rid of the dependence on the CNN backbone network, so much so that they cannot effectively extract global context information.

The accuracy and efficiency of semantic segmentation can be greatly increased by incorporating multi-scale feature fusion approaches with the aim of further enhancing the model's performance. Literature [37] points out that the Feature Pyramid Network (FPN) is an excellent strategy for combining multi-scale features, and for generating accurate predictions, a combination of high-resolution low-level features with high-level features with high semantic information is employed. By utilizing a pyramid pooling module to extract global contextual information, Zhao et al. [38] created a powerful multi-scale feature fusion Pyramid Scene Parsing Network (PSPNet) to increase segmentation accuracy. Chen et al. [29], [39], [40], [41] designed the DeepLab series of algorithms, among which DeepLabV3 [42] improved the hollow space pyramid pooling module by using hollow convolution using various expansion rates to record the target information at various scales and its related context. To capture the intended information as well as the contextual data to further mine the feature information at different scales and improve the segmentation effect. Chen et al. [41] created the kernel pyramid pooling (LKPP) module to extract different scale information to solve the feature extraction problem and optimize the boundaries with a new loss function. Wang et al. [43] used the dynamic multiscale dilation convolution to extract the features at different scales. Liu et al. [44] designed a dual-channel ASPP module for feature extraction. Wu et al. [45] proposed a multi-scale attention fusion module. Liu et al. [46] designed design context aggregation to capture multi-scale features through a pyramid network and region extractor. Cao et al. [47] captured important feature information by designing a multi-scale pyramid module. Although these methods can capture multi-scale features of remote sensing images, they do not adequately extract global information in terms of global information modeling.

Self-attention was initially used in machine translation [48] to capture remote features, and compared to the excessive computational effort of convolutional neural networks on remote sensing images, self-attention is able to ensure model expressiveness while reducing the computational effort. Combining the attention mechanism with convolutional neural networks can better utilize the model's productivity. Since then, the field of semantic segmentation has made extensive use of it. It has major benefits for building global contextual semantic information because it is based on modeling feature correlations in spatial and channel dimensions.

Recently, mechanisms of attention have also been incorporated into the study of semantic segmentation of remotely sensed images. Li et al. [49] designed the attention bilateral network ABCNet to model global remote dependencies with lightweight CNN spatial paths and contextual paths. A2FPN [50] enhances the model's multiscale features with

the attention aggregation module. DANet [51] proposes two attention modules modeling the feature dimensions of space and channel respectively, which improves the utilization of global information. Zhao et al. proposed a MANet [52] to obtain global context dependencies by extracting features through the attention module. Li et al. [53] pointed out a Multi-Level Attention Network (MAResU-Net) using the Linear Attention Mechanism (LAM) to establish long-range dependencies. Li et al. [54] proposed a fusing spatial and channel attention network SCAttNet, using lightweight spatial and channel attention for adaptive feature refinement, it even verified that the attention network is extremely valuable for improving accuracy. Bai et al. [55] proposed a dual-attention network DCAttNet with a DANet, which was designed to learn the spatial interdependence of features in the position attention module (PAM) and capture the channel interdependence in the channel attention module (CAM), enhanced the segmentation outcomes by simulating intricate contextual relationships with local characteristics. Wang et al. [56] suggested creating a bilateral awareness network (BANet) that combines contextual, global, and spatial information when constructing the attention feature aggregation module. Wang et al. [57] and others designed directed attention networks to learn orientation features and global semantic information of real objects. Lin et al. [58] designed spatial linear attention and channel linear attention mechanisms to capture global contextual remote dependencies. Song et al. [59] suggested the Full Attention Network (FLANet), which uses global contextual features to preserve spatial features when computing channel feature maps.

Inspired by the above literature, we propose a multi-scale full-attention fusion network (MFAFNet) based on an encoder-decoder structure for semantic segmentation of remote sensing images. In this paper, we design an explicit visual center (EVC) module, which captures local-global information. Lightweight pre-trained ResNet18 [49] was used as an encoder to extract local information. Then, a Full Attention Network (FANB) is designed to encode both channel and spatial attention simultaneously by introducing spatial interaction into the attention channel mechanism to obtain comprehensive worldwide contextual data. In order to appropriately fuse multi-scale variables and, thus, adequately capture contextual and spatial aspects, the study eventually adds the Gated Channel Attention Fusion (GCF) module into the decoder stage. Specifically, our principal contributions can be summed up as follows:

1) We propose a new semantic segmentation network, MFAFNet, to extract rich local-global contextual information and realize multi-scale feature fusion based on low computational cost.
2) In the feature encoding stage, we propose the EVC module and FANB network in conjunction with ResNet18 in the encoder to model global remote dependencies and capture local-global contextual information.
3) In order to effectively combine and extract high-level and low-level feature mappings with multi-scale features, we present the gated attention fusion module (GCF), which fuses information collected from several encoder and decoder stages together.

II. MATERIALS & METHODS

We will present the overall architecture and component modules of MFAFNet in this section. We first will give the general architecture of the network, and then explain the Explicit Visual Center module (EVC), the Fully Attentional Network block (FANB), and the Gate Channel Fusion Module (GCF) in detail.

A. OVERALL STRUCTURE

To cope with the semantic segmentation challenges of remotely sensed images in urban environments. We give a Multiscale Fully Attention Fusion Network for Remote Sensing Image Semantic Segmentation (MFAFNet). The MFAFNet architecture is displayed in Fig. 1, and the general structure consists of two parts: the encoder-decoder. For the encoder, to lower the computational cost of collecting multi-scale semantic information, we adopt ResNet18 with pre-trained weights as the feature extractor. Obtaining multi-scale semantic information by downsampling at a minimal computational cost. ResNet18 includes four stages, every stage with a scale factor of 2 to acquire four feature maps of varying sizes for the resolution: H/4×W/4, H/8×W/8, H/16×W/16 and H/32×W/32. The intricate spectral data and textural composition of urban remote sensing images, as well as irregular feature boundaries, these features put high demands on the extractor of features. Therefore, the detailed feature maps produced at the end of the encoder stage are used to further acquire global information through the EVC and FANB modules to make up for the insufficiency of local information. Furthermore, the feature maps generated in the remaining stages be employed for acquiring rich global information at multiple scales through the FANB module. The four feature mappings produced by the encoder are fused with the feature mapping GCF produced by the decoder during the decoder stage. By using a 1×1 convolution operation, the fusion process is unified to a 64-channel dimension. Specifically, the high-level semantic characteristics produced by the decoder are merged and weighted with the semantic features produced by FANB. The weighted summing operation adaptively adjusts the magnitude of the weights according to the contributions of these two features, which can be expressed as:

$$FU = \alpha FANB + (1 - \alpha) \cdot GCF \tag{1}$$

where FU denotes the fused feature, FANB stands for the feature Generated by the FANB Module, and GCF represents the feature produced by the Gate Channel Fusion Module.

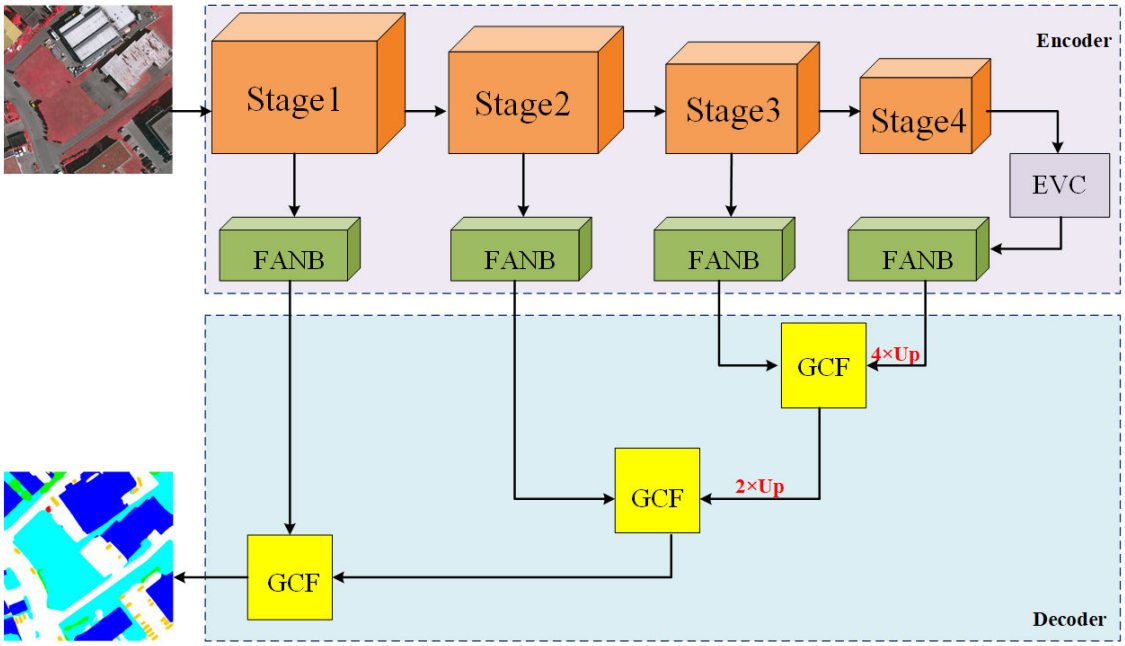

FIGURE 1. Overall network structure of the MFAFNet.

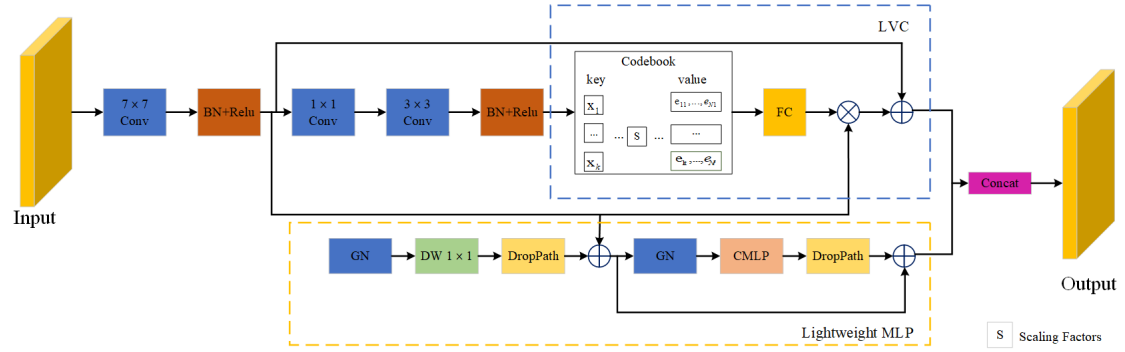

FIGURE 2. Illustration of the EVC module.

B. EVC MODULE

While current methods primarily concentrate on the interactions between features at different layers, they do not address the conditioning of features within the same layer, even though studies have demonstrated the advantages of using intra-layer feature rules in visual identification tasks. In our work, inspired by the work of [60], we propose Explicit Visual Centre EVC. specifically, we adopt a parallel learnable center method to collect local corner regions of high-level features produced by ResNet18, and a lightweight MLP to collect global contextual information. As Fig. 2 shows, our proposed EVC mainly contains two parallel connected blocks: the lightweight MLP and the learnable visual center LVC, where the lightweight MLP gets the global information of the high-level feature X_4 while reducing the computational cost, but to minimize the loss of local information, we use the LVC learnable visual center mechanism on the high-level feature X_4 generated by ResNet18 to preserve local corner regions by aggregating localized features of the region within the layer. The output feature maps of these two blocks are connected along the channel dimensions as input to FANB. The above process can be represented as:

$$F_X = cat(MLP(X_4), LVC(X_4)) \tag{2}$$

where F_X is the EVC output and cat$(\cdot)$ are denoted as two blocks of feature maps concatenated localized features of the region within the layer, and are denoted as the feature maps output using the lightweight MLP and the learnable mechanism LVC respectively. They are described in detail next.

The lightweight MLP in the EVC module is mainly to record high-level features' global remote dependencies, which is mainly composed of two sections: the channel MLP-based convolution and the depth convolution-based module [61], where the depth convolution module is the input to the channel MLP module. Strengthening the robustness and generality of the features, we implement the channel scaling operation [62] and the DropPath operation at the end. Specifically, X_4 is the first input to the deep convolution, which is processed by group norming. This is followed

by channel scaling and DropPath, and then finally the residual concatenation of X_4 is performed. Compared with the traditional spatial convolution, the deep convolution used in this paper not only lowers computing costs but also enhances the capacity for feature representation. The procedures mentioned above can be expressed as:

$$X_D = DWConv(GN(X_4)) \tag{3}$$

$$\tilde{X}_D = Drop(CS(X_D)) + X_4 \tag{4}$$

where X_D represents the output, GN(·) is the group normalization and DWConv(·) is a depthwise convolution with the kernel size 1×1. Channel scaling (CS) and Drop (DropPath) denote the output of the depth convolution-based module, respectively. Afterward, we feed the resulting features into the channel-based MLP module.

Specifically, the X_D is first transported into the group normalization, then the channel MLP [63] is put into practice in the generated features, similar to the operation of the depth-based convolution module, followed by channel scaling and DropPath, and finally, the remaining connections of the X_D are added again to prevent overfitting. Compared with spatial MLP, Channel MLP satisfies the needs of general vision applications while also efficiently reducing computing complexity [64]. The procedure mentioned above can be written like this:

$$X_M = CMLP(GN(\tilde{X}_D)) \tag{5}$$

$$\tilde{X}_M = Drop(CS(X_M)) + \tilde{X}_D \tag{6}$$

where CMLP(·) is the channel MLP. X_M is the output of the channel MLP, $\tilde{X}_M$ denotes the final output.

The encoder LVC has an internal dictionary, which can preserve local corner regions and reduce the loss of local information. LVC consists of two parts: Code-book: $E = \{e_1, e_2, \ldots. e_t\}$, and the scaling factor $S = \{s_1, s_2, \ldots s_t\}$ are the learnable visual centers, where t represents the total number of visual centers. The feature X_4 after convolution, after 1×1 and 3×3 convolutions, is transported into the codebook E. In the codebook, we use a set of scaling factors S to map make X_K and E_t sequentially to the corresponding positional information, which is represented as follows:

$$X_4 = Relu(BN(Conv_{7\times7}(X_4))) \tag{7}$$

$$\tilde{X}_4 = Relu(BN(Conv_{3\times3}(Conv_{1\times1}(X_4)))) \tag{8}$$

$$C = \sum_{t=1}^{t} \phi\Big(\sum_{i=1}^{N} \frac{e^{-s_t\|\tilde{x}_k - e_t\|}}{\sum_{j=1}^{t} e^{-s_t^{\|\tilde{x}_k - e_t\|^2}}}(\tilde{x}_k - e_t)\Big) \tag{9}$$

where $\tilde{x}_k$ is the k-th pixel point, e_t is the t-th learnable visual code-word, $x_k - e_t$ is the information about each pixel position relative to a codeword is a t-th scaling factor, and t represents the total number of visual centers. N=H×W is the total spatial number of the input features, where H and W denote the feature map spatial size in height and width, respectively. ϕ is used to calculate the entire codebook. C denotes the whole codebook. Afterward, the output of the feature through the codebook and scaling factor are then multiplied and then summed with the initial features $\tilde{X}_4$ through the fully connected layer to get the final result of the LVC output. The above processes are expressed as:

$$X_{LVC} = \tilde{X}_4 \oplus \tilde{X}_4 \otimes FC(C) \tag{10}$$

where FC(·) is the fully connection layer, $\oplus$ s channel-wise multiplication, and $\otimes$ is the channel-wise addition. X_{LVC} denotes the final output.

C. FANB MODULE

Although non-local self-attention methods have been effective in capturing remote dependencies for semantic segmentation in recent years, these methods usually compress the spatial dimension or compress the similarity graph of the channel dimension to express their feature relationships. This approach tends to compress feature dependencies in other dimensions, leading to poor segmentation results for small/thin categories or inconsistent segmentation within large objects. To attempt to tackle this problem, we suggest a new approach, Full Attention Networks (FANB). Specifically, to avoid adding extra computational cost, spatial and channel attention are encoded in a single similarity graph adding spatial interactions to the channel attention mechanism and employing global average pooling as a global context prior.

According to Figure 3, by taking the feature maps generated by the ResNet18 or EVC module we input them into the FANB module as its initial features, where C is the number of channels and H and W represent the height and width of the feature maps respectively. Firstly, we input the feature map $X_F \in \mathbb{R}^{C\times H\times W}$ into Q, K, and V. Taking Q as an example first, to acquire a comprehensive, contextualized global prior, we input theX_F into two parallel global average poolings and choose two pooling windows that are unequal in height and width, $H \times 1$ and $1 \times W$ pooling windows. Then to ensure that each spatial location is connected to the corresponding global prior with the same horizontal or vertical coordinates, the choice is made to keep the one-dimensional length constant and output the features. Afterward, the global features $\tilde{Q}_h \in \mathbb{R}^{C\times H\times W}$, $\tilde{Q}_w \in \mathbb{R}^{C\times H\times W}$ in the vertical and horizontal directions are then obtained by repeating Q_H and Q_W, and the final feature map $Q \in \mathbb{R}^{(H+W)\times C\times N}$ is obtained by matrix multiplication of $\tilde{Q}_H$ and $\tilde{Q}_W$. In the same way, as for Q generation, generate $K \in \mathbb{R}^{(H+W)\times C\times N}$ and $V \in \mathbb{R}^{(H+W)\times C\times N}$.

The feature response from the larger context needs to be approved for each specific spatial location before the corresponding column node, we capture full attention through an affine operation as follows:

$$A_{i,j} = \frac{exp(Q_i \cdot K_j)}{\sum_{i=1}^{C} exp(Q_i \cdot K_j)} \tag{11}$$

where $A_{i,j}$ indicates the correlation dependency between the ith and jth channel at a particular spatial location.

Finally, we perform matrix multiplication of A and V obtained through affine operation to update each channel mapping, and then divide the obtained channel mapping into two groups, each of size $\mathbb{R}^{C\times H\times W}$, and sum t the two groups were gaining the information in the global remote context. The obtained context information is multiplied by the scale parameter γ, and the final feature map $F \in \mathbb{R}^{C\times H\times W}$ is obtained by summarizing each ingredient separately and 3×3 convolution with the initial feature map XF as shown below:

$$\tilde{F} = \sum_{i=1}^{C} A_{i,j} \cdot V_j + X_F \tag{12}$$

$$F = Conv_{3\times 3}(\tilde{F}) \tag{13}$$

where F represents the final output.

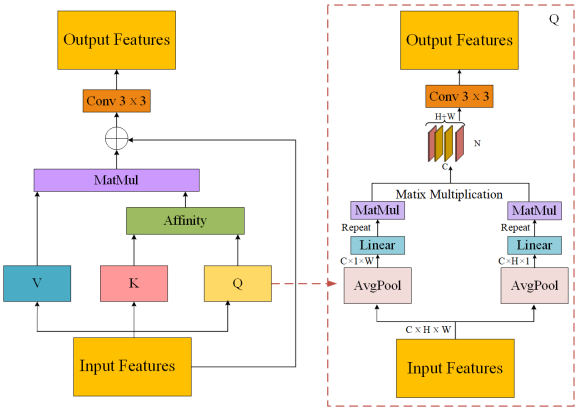

FIGURE 3. Illustration of the FANB module.

D. GCF MODULE

Because of the wide variations in target sizes in high-resolution remote sensing images, it is not viable to efficiently use the multi-scale features of the feature map when semantic segmenting remote sensing images. Therefore, in order to more effectively integrate the multi-scale features, we suggest the GCF module, and unlike previous fusion methods, our proposed method utilizes a gating mechanism to select useful information among a large amount of invalid information, which improves the efficiency and accuracy at the same time, as shown in Fig. 4.

In GCF, there are two input feature maps, F_{high} and F_{low}, in which the low-dimensional feature map's resolution is half that of the high-dimensional feature map, and such a setup restricts the propagation of information between feature maps, which is advantageous for the merging of information. Firstly, the feature map F_{high} is convolved 3×3, in order for the two feature maps' channel dimensions to be consistent, and then the optimized high-dimensional and low-dimensional feature maps of F_{high} and F_{low} are aggregated and optimized by concat operation, and then the 1×1 convolution sum and a Sigmoid activation function. To better capture data from multiple scales in the process of fusion and to improve the segmentation

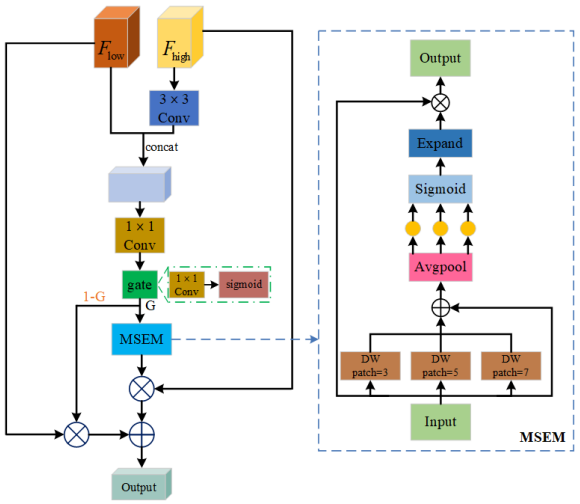

FIGURE 4. Illustration of the GCF module.

accuracy, we designed a multi-sale separation enhancement module (MSEM) in GCF. Specifically, since different sizes of receptive fields imply different abilities to be able to record distant dependencies, we first use the extended convolution with expansion factors [3,5,7] to fully utilize the receptive fields of the feature maps, and three different extended convolution rates are used to capture multiscale information and different ranges of dependencies. Then the resulting feature maps are fed into the adaptive pooling and the original features are residually concatenated and adaptively pooled to prevent overfitting. Then, the obtained feature map is input into the adaptive pooling and the original features are connected by residuals to prevent overfitting and adaptive pooling, the pooled results are weighted and operated with the pooled results after extended convolution. Finally, the attention weights are generated by linear transformation and sigmoid activation function, and the generated weights are multiplied with the original features element by element to get the final output, which further improves the ability to perceive the important information in the fusion process. Next, the gating value G is transported to output G_M in the multiscale self-attention module MSEM to enhance the gating representation and obtain rich multiscale information. The feature map F_{high} is multiplied element by element with the gating value G_M, F_{low}, and 1-G_M, respectively, to obtain the feature maps F_{HM} and F_{LM}, which can dynamically adjust the weight distribution between them to better capture the details that are difficult to be captured or the classifications that are neglected in the segmentation process. Gating operations are eventually followed by a weighted summing on the feature maps to generate the fused processed feature maps.

III. RESULTS

A. DATASETS

The ISPRS Potsdam, ISPRS Vaihingen, and LoveDA datasets are publicly accessible datasets that can be used for model training and testing.

The Vaihingen dataset, which was gathered in the German region of Vaihingen and supplied by the ISPRS Working Group III/4 under the framework of the "ISPRS Urban Classification and 3D Building Reconstruction Test Project," is the dataset utilized in this work. The dataset is made up of 33 orthophotos, each measuring an average of 2494 × 2064 with a ground sampling distance (GSD) of 9 cm. Each image has three bands: red, green, and near-infrared, along with the associated normalized DSM and digital surface model (DSM). It consists of six categories: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. In the experiment. We utilized ID: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38 for testing, and the remaining 16 images for training. Each image and label were manually cropped to a size of 1024 × 1024 pixels.

There are 38 orthophotos in the Potsdam collection, with a ground resolution (GSD) of 5 cm and a pixel size of 6000 × 6000. Along with the matching Digital Surface Model (DSM) and Normalized Difference of the Digital Surface Model (NDSM), each image includes red, green, blue, and near-infrared bands. This dataset, like the Vaihingen dataset, is divided into six categories: automobiles, low vegetation, impermeable surfaces, buildings, and clutter/background. We utilized ID: 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 for testing, and the remaining 23 images, except image 7_10 with error annotations for training. In the experiments, we only used the red, green, and blue bands, and we reduced the original image tiles to 1024 × 1024 pixels.

LoveDA dataset is a surface cover classification dataset proposed by the RSIDEA team, which contains 5987 0.3 m high-resolution images originating from the cities of Wuhan, Nanjing, and Changzhou. The resolution of each image is 1024 × 1024 pixels and the land cover categories include building, road, water, wasteland, forest, agriculture, and background, of which the number of images used for training is 2522, the number of validations is 1669, and the number of test is 1796. The dataset is challenging considering the diversity and complexity of land cover types and background samples in different scenes.

B. EXPERIMENTAL SETTINGS

All of the experiments were conducted using the PyTorch framework on a Linux operating system and a single NVIDIA 3090 GPU as the hardware foundation. We chose the AdamW optimizer for the model training procedure to guarantee the equity of the comparison findings. The initial learning rate was 6e-4, the weight decay value was set to 0.01 and the learning rate was dynamically adjusted using the cosine approach.

When training on the Vaihingen and Potsdam datasets, we employed a random cropping strategy, randomly cropping the images into 512 × 512 patches, to prevent overfitting. We employed several data improvement strategies during the training phase, such as random rotation, random vertical flip, random horizontal flip, and random scaling ([0.5,0.75,1.0,1.25,1.5]). In the meantime, we established a maximum of 110 training rounds and split the training data into batches of eight samples each. We employed the test time augmentation (TTA) technique in addition to multi-scale [0.5,0.75,1,1.25,1.5] throughout the testing phase.

C. EVALUATION INDICATORS

To comprehensively measure the effectiveness of the model we have proposed, we used three evaluation metrics, including overall accuracy (OA), the mean intersection over union (mIoU), and the mean F1 score (mF1). Based on the accumulated confusion matrix, OA, mIoU, and mF1 are calculated as follows:

$$OA = \frac{\sum_{K=1}^{N} TP_K}{\sum_{K=1}^{N} TP_K + FP_K + TN_K + FN_K} \tag{14}$$

$$mLOU = \frac{1}{N}\sum_{K=1}^{N} \frac{TP_K}{TP_K + FP_K + TN_K + FN_K} \tag{15}$$

$$precision = \frac{1}{N}\sum_{K=1}^{N} \frac{TP_K}{TP_K + FP_K} \tag{16}$$

$$recall = \frac{1}{N}\sum_{K=1}^{N} \frac{TP_K}{TP_K + FN_K} \tag{17}$$

$$F1 = \frac{1}{N}\frac{precision \times recall}{precision + recall} \tag{18}$$

where TP_K, FP_K, and TN_K stand for the corresponding true positives, false positives, true negatives, and false negatives, for objects indexed as class k.

In addition, we use the number of floating-point operations per second (Flops) and the number of parameters to evaluate the complexity of the model.

D. PERFORMANCE COMPARISON

1) COMPARISON WITH STATE-OF-THE-ART METHODS ON ISPRS VAIHINGEN

The experimental findings using the ISPRS Vaihingen dataset, as indicated in Table 1, provide a comparison of the various approaches. But as with earlier tests, there is no reporting of the backdrop and clutter accuracy, The best MeanF1/mloU/OA was achieved with our proposed MFAFNet method. When compared to the suboptimal technique, there was an improvement of 0.94% in MeanF1, 1.57% in mloU, and 0.85% in OA. Additionally, our approach received the best ratings across all categories. It is noteworthy that our method obtained an 89.23% F1 score on the "car" class, outperforming other networks by more than 0.67%. We show the results of the visualization of each method on the test set in Fig. 5. We show the visualization of each method on the test in Fig. 5, and based on the segmentation results, as we can see, our approach performs better than other methods.

TABLE 1. Quantitative comparison with state-of-the-art models on the ISPRS VAIHINGEN dataset. The best values in the columns are in bold. All scores are expressed as percentages (%). Where the metric for all categories is F1 scores.

Method	Backbone	Imp.surf	Building	Lowveg	Tree	Car	MeanF1	OA	mIoU
DABNet	ResNet18	95.97	92.98	82.51	88.86	82.77	88.62	91.75	79.98
U-Net	ResNet18	95.23	91.34	80.54	87.74	79.38	86.85	90.56	77.27
ERFNet	ResNet18	95.66	92.78	81.99	88.39	77.24	87.21	91.35	77.96
A^2FPN	ResNet18	96.56	95.40	83.69	89.51	87.68	90.57	92.89	83.10
BiSeNet	ResNet18	96.50	94.86	83.35	89.09	82.13	89.18	92.56	80.98
MANet	ResNet18	96.51	95.21	83.16	89.45	85.91	90.05	92.74	82.30
MAResU-Net	ResNet18	96.64	95.44	83.91	89.73	86.67	90.48	92.98	82.98
ABCNet	ResNet18	95.70	92.18	81.56	88.56	79.47	73.60	91.19	76.76
BANet	ResNet18	96.55	95.17	83.47	89.46	88.56	90.64	92.80	83.23
UNetFormer	ResNet18	96.64	95.37	83.56	89.61	86.68	90.37	92.92	82.82
MFAFNet	ResNet18	**97.02**	**95.81**	**85.50**	**90.50**	**89.23**	**91.61**	**93.65**	**84.81**

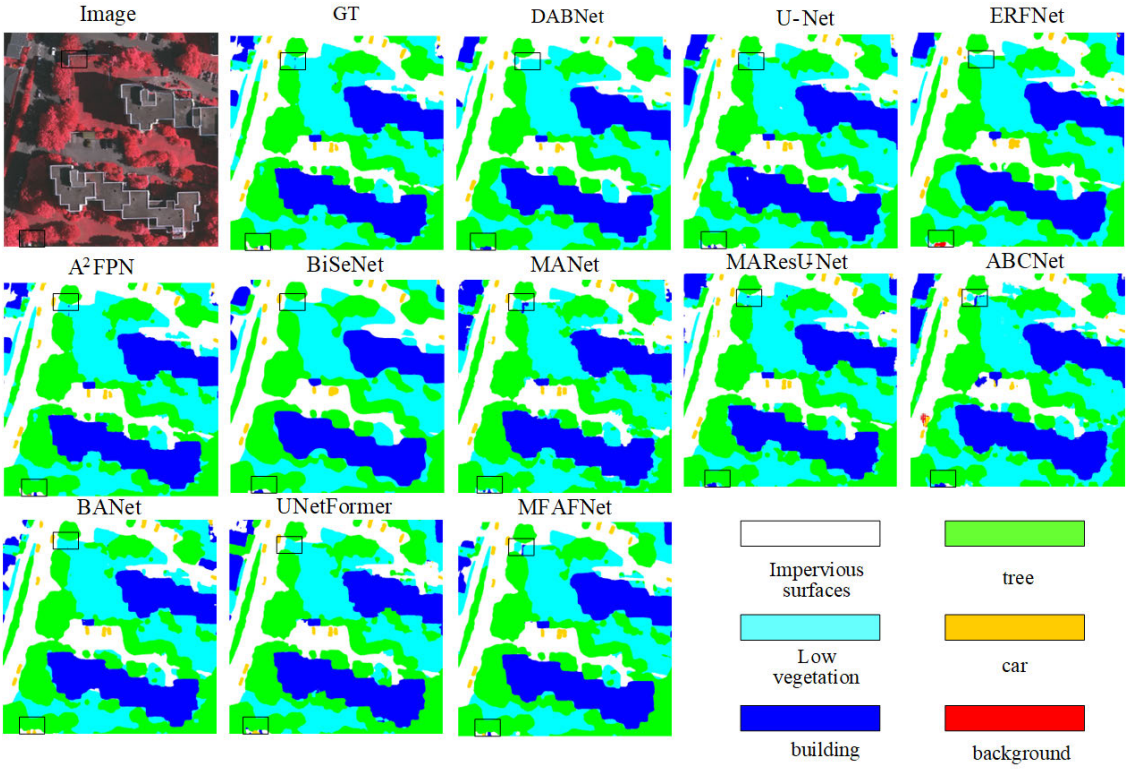

FIGURE 5. Qualitative comparison of the visualization results of our method with other methods on the Vaihingen dataset.

2) COMPARISON WITH STATE-OF-THE-ART METHODS ON ISPRS POTSDAM

Using the ISPRS Potsdam dataset, we carried out a comparative experiment of the widely used approaches in order to further confirm the generalizability of the MFAFNet model. Table 2 presents a list of the specific comparative results. There is no report about clutter or background accuracy in the Potsdam test set, which is similar to the Vaihingen dataset. Table 2 illustrates that our MFAFNet model performs satisfactorily on the Potsdam test set, achieving 92.51% for F1, 86.27% for mIoU, and 91.30% for OA. Compared to the suboptimal method, MeanF1 improved by 0.59%, mIoU by 1%, and OA by 0.56%. And compared to other methods, our method reached the highest scores in the categories. For example, our method obtained a 95.93% F1 score on the "car" class, outperforming other networks by more than 0.41%. In addition, we display the corresponding visualization results in Fig. 6 to compare and illustrate the differences between our suggested method and other commonly used methods. It is evident from looking at Fig. 6 that the segmentation results using the method employed in this paper have more distinct edge contours and more information.

3) COMPARISON WITH STATE-OF-THE-ART METHODS ON LoveDA

To further evaluate the effectiveness of the proposed approach, we conducted comparative experiments on the widely used methods in the LoveDA dataset. The specific comparison results have been presented in Table 3. The experiments show that our proposed MFAFNet model achieves the best results on the LoveDA test set with an mIoU

TABLE 2. Quantitative comparison with state-of-the-art models on the ISPRS POTSDAM dataset. The best values in the columns are in bold. All scores are expressed as percentages (%). Where the metric for all categories is F1 scores.

Method	Backbone	Imp.surf	Building	Lowveg	Tree	Car	MeanF1	OA	mIoU
DABNet	ResNet18	90.28	92.70	82.46	81.66	92.24	87.87	86.45	78.68
ERFNet	ResNet18	89.79	91.90	78.15	77.06	91.61	85.70	84.20	75.56
A^2FPN	ResNet18	93.46	95.86	86.76	88.02	95.49	91.92	90.74	85.27
BiSeNet	ResNet18	93.09	95.68	86.29	88.24	94.96	91.65	90.47	84.81
MANet	ResNet18	92.89	95.77	86.23	87.51	95.30	91.54	90.31	84.65
MAResU-Net	ResNet18	93.42	95.80	86.57	87.84	95.41	91.81	90.56	85.09
ABCNet	ResNet18	89.96	92.21	80.58	79.09	89.79	86.32	85.19	76.33
BANet	ResNet18	93.43	96.00	86.82	88.41	94.90	91.91	90.82	83.24
UNetFormer	ResNet18	92.94	95.18	86.33	87.50	95.52	91.49	90.15	84.55
MFAFNet	ResNet18	**93.74**	**96.08**	**87.65**	**89.18**	**95.93**	**92.51**	**91.30**	**86.27**

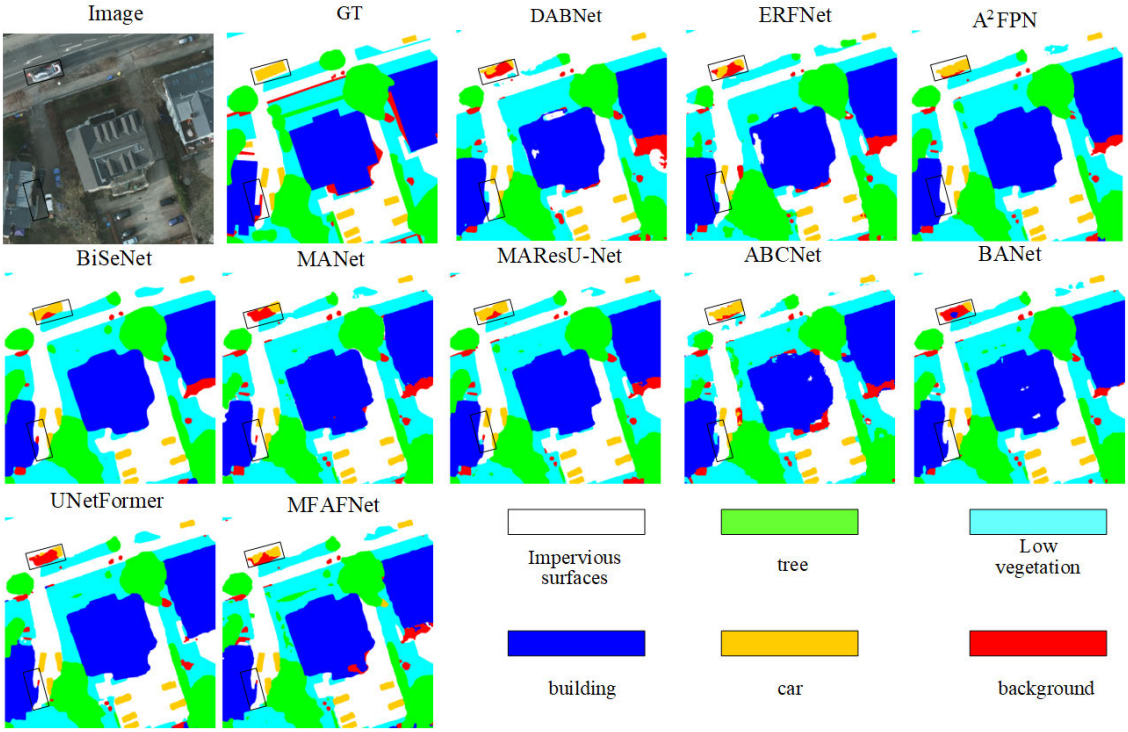

FIGURE 6. Qualitative comparison of the visualization results of our method with other methods on the Potsdam dataset.

of 52.11%. Moreover, the visualization results obtained by the corresponding method are also shown in Fig. 7. From the visualization results, it can be clearly observed that in terms of edge segmentation and target object segmentation, our method performs more smoothly and accurately compared to other methods.

E. ABLATION EXPERIMENTS

1) EFFECT OF EACH COMPONENT OF MFAFNet

We performed a series of ablation experiments on the Vaihingen dataset to assess the performance of each component in MFAFNet. We obtained the results displayed in Table 4 by individually assessing each module's operation. We build the benchmark model in the baseline model using ResNet18 with the U-Net as the backbone network. Only local context information in the decoder is modeled. The Fully Attentional Network block (FANB): Four Fully Attention Network blocks were incorporated into the baseline to build the Baseline+FANB. Performance improved for all categories with the addition of the FANB module, as Table 4 demonstrates, Baseline+FANB achieves mIoU growth of 6.36% on the Vaihingen test sets, which proves the positive effects of the FANB module. The Explicit Visual Center module (EVC): We inserted the Explicit Visual Center module into Baseline +FANB as the decoder. As illustrated in Table 4, the introduction of the EVC module improves the mIoU of the network on the Vaihingen dataset by 0.27%, demonstrating the EVC module's efficacy. The Gate Channel Fusion Module (GCF): The GCF module was made available to construct the full MFAFNet network. As shown in Table 4, the introduction of the GCF module improves the mIoU of the network on the Vaihingen dataset by 0.1% and obtained the highest MeanF1, mIoU.

As seen in Fig. 8, we also display the results for each module. When all modules are added, the network can efficiently extract global-local contextual information and multi-scale information to more closely match GT values.

TABLE 3. Quantitative comparison with state-of-the-art models on the LoveDA dataset. The best values in the columns are in bold. All scores are expressed as percentages (%).

Method	Backbone	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU
DABNet	ResNet18	38.55	50.13	53.71	74.41	15.84	41.03	52.64	46.62
ERFNet	ResNet18	37.17	49.77	45.59	72.39	16.70	37.83	47.21	43.81
A^2FPN	ResNet18	45.92	57.46	**55.66**	79.11	**18.45**	43.86	62.56	51.86
BiSeNet	ResNet18	45.73	57.19	55.35	78.46	16.89	45.31	63.72	51.81
MANet	ResNet18	**50.88**	43.81	56.16	52.70	18.00	45.89	60.87	50.88
MAResU-Net	ResNet18	44.77	55.37	53.33	78.20	15.34	45.74	60.63	50.48
ABCNet	ResNet18	36.55	40.07	40.02	70.59	12.71	38.80	54.88	41.95
BANet	ResNet18	46.02	**58.17**	51.03	78.47	14.93	45.54	62.79	51.00
UNetFormer	ResNet18	44.52	57.60	54.73	78.56	16.41	46.38	62.06	51.47
MFAFNet	ResNet18	46.43	57.74	55.15	**79.83**	15.28	**47.47**	**62.88**	**52.11**

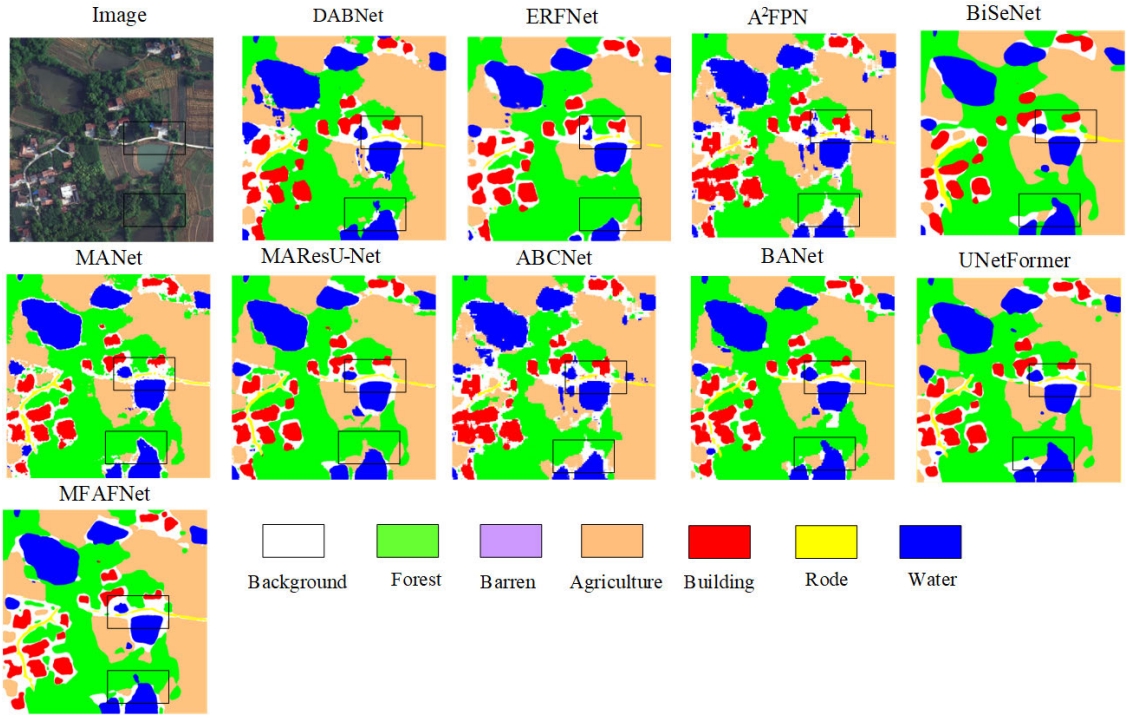

FIGURE 7. Qualitative comparison of the visualization results of our method with other methods on the LoveDA dataset.

TABLE 4. Ablation study of each component of the MFAFNET. The best values in the columns are in bold. All scores are expressed as percentages (%), where the metric for all categories is F1 scores.

Dataset	Method	Imp.surf	Building	Lowveg	Tree	Car	MeanF1	OA	mIoU	Params (Mb)	FLOPs (Gbps)
Vaihingen	Baseline	95.80	94.19	83.00	88.67	74.08	87.15	91.92	78.08	**11.51**	**32.69**
	Baseline+FANB	96.89	95.74	84.90	90.34	89.05	91.38	93.46	84.44	11.91	34.81
	Baseline+FANB+EVC	96.97	**95.88**	85.29	**90.52**	89.08	91.55	93.60	84.71	13.29	35.93
	Baseline+FANB+EVC+GCF	**97.02**	95.81	**85.50**	90.50	**89.23**	**91.61**	**93.65**	**84.81**	15.00	41.55

2) EFFECT OF THE FANB MODULE

The functioning of spatial attention and channel attention is examined in this section. Table 5 displays the results of the experiment. There was a 0.7% decrease in mIoU, a 0.98 M reduction in parameters, and a 5.89 G decrease in FLOPs when merely channel attention was added. Table 5 lists the results. When only spatial attention is added, the mIoU decreases by 1.17%, the parameters are reduced by 1M, and the FLOPs are reduced by 5.9 G. This illustrates the absence of both channel and spatial attention without excessive computational cost overheads.

3) EFFECT OF EVC MODULE

The EVC module comprises two parts. connected in parallel: lightweight MLP(LMP) and learnable vision center (LVC). The experimental results are displayed in Table 6. We find that the average intersection ratio mIoU decreases by 0.12% when solely LVC is implemented. We disregarded the parameter adjustments and FLOP operations at this point. On the other hand, since LVC is not used when simply LMP is added, the mIoU drops by 0.46%. This leads one to conclude that when both operations are applied simultaneously, the best segmentation accuracy can be guaranteed.

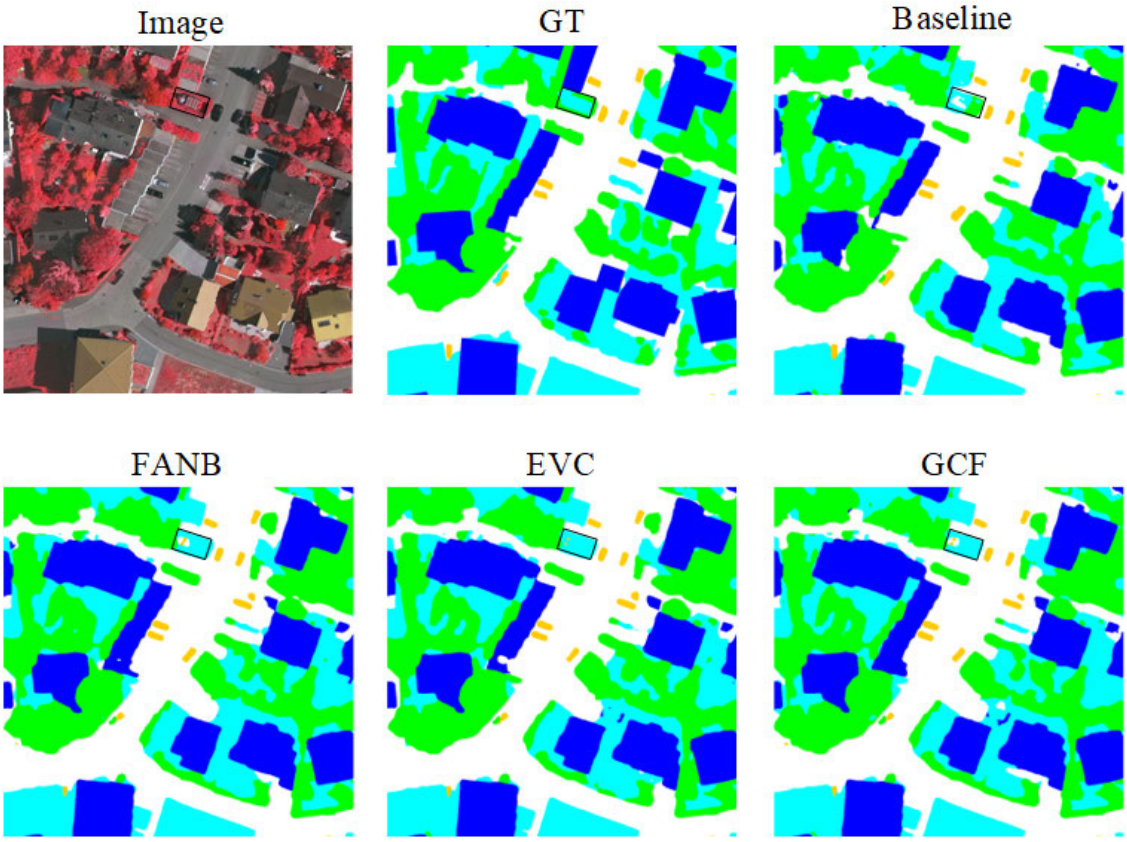

FIGURE 8. Visualization of ablation experiments on the Vaihingen dataset.

TABLE 5. Results of ablation experiments on the VAIHINGEN dataset for the FANB module.

Method	Channel	Spatial	Params (Mb)	FLOPs (Gbps)	mIoU
FANB	√		14.02	35.66	84.11
FANB		√	14.00	35.65	83.64
FANB	√	√	15.00	41.55	84.81

TABLE 6. Results of ablation experiment on the VAIHINGEN dataset for the EVC module.

Method	LVC	LMP	Params (Mb)	FLOPs (Gbps)	mIoU
EVC	√		14.95	41.51	84.69
EVC		√	14.97	41.53	84.35
EVC	√	√	15.00	41.55	84.81

4) EFFECT OF GCF MODULE

In the GCF Module, (separated and enhanced attention module) MSEM Attention Module has a significant impact, in which MSEM is related to the size of the patch size. Table 7 shows the experimental data. When the patch size [1,3,5], the mIoU is reduced by 0.45%. When the patch size [2,4,6], the mIoU is reduced by 0.18%. It is obvious from the experiment that when the patch size is [3,5,7], the effect is better and the FLOPs reach 41.55G.

TABLE 7. Results of ablation experiment on the VAIHINGEN dataset for the GCF module.

Method		Params (Mb)	FLOPs (Gbps)	mIoU
GCF	patch size=[1,3,5]	14.41	41.93	84.36
GCF	patch size=[2,4,6]	14.67	41.62	84.66
GCF	patch size=[3,5,7]	15.00	41.55	84.81

TABLE 8. Comparison of the complexities of our method with other methods. The best values in the columns are in bold.

Method	Params (Mb)	FLOPs (Gbps)	mIoU
DABNet	**0.75**	**16.93**	79.98
U-Net	2.06	47.40	77.96
ERFNet	7.78	545.00	77.27
A^2FPN	22.82	135.87	83.10
BiSeNet	13.42	48.96	80.98
MANet	12.0	71.02	82.30
MAResU-Net	16.17	81.33	82.98
ABCNet	14.0	50.01	76.76
BANet	12.76	42.60	83.23
UNetFormer	11.68	37.57	82.82
MFAFNet	15.0	41.55	**84.81**

F. MODEL COMPLEXITIES

Table 8 of the results presented in this research lists the parameter sizes and computational complexity for each approach under identical operating conditions. Additionally,

as the table illustrates, MFAFNet shows excellent results compared to other methods and is less computationally expensive compared to MAResU-Net and A^{2}FPN.

IV. CONCLUSION

In this paper, we propose a Multiscale Fully Attention Fusion Network for Re-mote Sensing Image Semantic Segmentation (MFAFNet). MFAFNet uses lightweight ResNet18 for down-sampling in the encoding stage, reducing the computational cost while extracting local information. After that, the explicit visual center module EVC and the full attention network FANB are used to extract rich global context information to improve the segmentation accuracy of the model. Finally, the gated channel attention fusion module GCF is used in the decoder stage, which not only enhances the channel interaction but also realizes the fusion of low and high-dimensional features, capturing multi-scale features. The validity of the MFAFNet method proposed in this study was confirmed by a combination of comparative tests and ablation experiments using the Vaihingen, Potsdam, and LoveDA datasets. The next research will focus on balancing efficiency and segmentation accuracy, as well as further optimizing the network structure.

REFERENCES

[1] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114417.

[2] H. Wuit Yee Kyaw, A. Chatzidimitriou, J. Hellwig, M. Bühler, J. Hawlik, and M. Herrmann, "Multifactorial evaluation of spatial suitability and economic viability of light green bridges using remote sensing data and spatial urban planning criteria," *Remote Sens.*, vol. 15, no. 3, p. 753, Jan. 2023.

[3] H. Zhang and S. Liu, "Double-branch multi-scale contextual network: A model for multi-scale street tree segmentation in high-resolution remote sensing images," *Sensors*, vol. 24, no. 4, p. 1110, Feb. 2024.

[4] B. Gui, A. Bhardwaj, and L. Sam, "Evaluating the efficacy of segment anything model for delineating agriculture and urban green spaces in multiresolution aerial and spaceborne remote sensing images," *Remote Sens.*, vol. 16, no. 2, p. 414, Jan. 2024.

[5] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sens. Environ.*, vol. 236, Jan. 2020, Art. no. 111402.

[6] W. Hou, Y. Wang, J. Su, Y. Hou, M. Zhang, and Y. Shang, "Multi-scale bilateral spatial direction-aware network for cropland extraction based on remote sensing images," *IEEE Access*, vol. 11, pp. 109997–110009, 2023.

[7] J. Chen et al., "Efficient seismic data denoising via deep learning with improved MCA-SCUNet," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024, Art. no. 5903614.

[8] G. Chen, J. Chen, K. Jensen, C. Li, S. Chen, H. Wang, J. Li, Y. Qi, and X. Huang, "Joint data and model-driven simultaneous inversion of velocity and density," *Geophys. J. Int.*, vol. 237, no. 3, pp. 1674–1698, Apr. 2024.

[9] R. Li, S. Zheng, C. Duan, L. 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.

[10] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.

[11] S. Subudhi, R. N. Patro, P. K. Biswal, and F. Dell'Acqua, "A survey on superpixel segmentation as a preprocessing step in hyperspectral image analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5015–5035, 2021.

[12] L. Wei, X. Liu, X. Li, and H. Gao, "System dynamics simulation and regulation of human-water system coevolution in Northwest China," *Frontiers Ecol. Evol.*, vol. 10, Jan. 2023, Art. no. 1106998.

[13] Q. Gao, D. Liu, W. Zhang, and Y. Liu, "Deep learning-based key indicator estimation in rivers by leveraging remote sensing image analysis," *IEEE Access*, vol. 12, pp. 72277–72287, 2024.

[14] S. Tian, Y. Zhong, Z. Zheng, A. Ma, X. Tan, and L. Zhang, "Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application," *ISPRS J. Photogramm. Remote Sens.*, vol. 193, pp. 164–186, Nov. 2022.

[15] L. S. Davis, A. Rosenfeld, and J. S. Weszka, "Region extraction by averaging and thresholding," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, no. 3, pp. 383–388, Mar. 1975.

[16] M. Özden and E. Polat, "Image segmentation using color and texture features," in *Proc. 13th Eur. Signal Process. Conf.*, Sep. 2005, pp. 1–4.

[17] N. Senthilkumaran and R. Rajesh, "Image segmentation–A survey of soft computing approaches," in *Proc. Int. Conf. Adv. Recent Technol. Commun. Comput.*, Oct. 2009, pp. 844–846.

[18] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1–9.

[19] S. Nowozin, "Structured learning and prediction in computer vision," *Found. Trends Comput. Graph. Vis.*, vol. 6, nos. 3–4, pp. 185–365, 2010.

[20] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[21] W. Han, R. Feng, L. Wang, and L. Gao, "Adaptive spatial-scale-aware deep convolutional neural network for high-resolution remote sensing imagery scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 4736–4739.

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

[23] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.

[24] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[26] G. Chen, C. He, T. Wang, K. Zhu, P. Liao, and X. Zhang, "A superpixel-guided unsupervised fast semantic segmentation method of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[28] 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. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[29] 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. (ECCV)*, 2018, pp. 801–818.

[30] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[31] M. Cui, K. Li, J. Chen, and W. Yu, "CM-Unet: A novel remote sensing image segmentation method based on improved U-Net," *IEEE Access*, vol. 11, pp. 56994–57005, 2023.

[32] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022, Art. no. 4408820.

[33] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[34] 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. (ECCV)*, 2018, pp. 325–341.

[35] G. Li, I. Yun, J. Kim, and J. Kim, "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," 2019, *arXiv:1907.11357*.
[36] J. Hou, Z. Guo, Y. Feng, Y. Wu, and W. Diao, "SPANet: Spatial adaptive convolution based content-aware network for aerial image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2192–2204, 2023.
[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.
[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[41] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[42] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 15–28, Dec. 2020.
[43] Z. Wang, S. Zhang, L. Gross, C. Zhang, and B. Wang, "Fused adaptive receptive field mechanism and dynamic multiscale dilated convolution for side-scan sonar image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022, Art. no. 511681.
[44] W. Liu, Y. Zhang, J. Yan, Y. Zou, and Z. Cui, "Semantic segmentation network of remote sensing images with dynamic loss fusion strategy," *IEEE Access*, vol. 9, pp. 70406–70418, 2021.
[45] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, Art. no. 2004612.
[46] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, and C. W. Chen, "Learning to aggregate multi-scale context for instance segmentation in remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 23, 2024, doi: 10.1109/TNNLS.2023.3336563.
[47] D. Cao, J.-N. Cao, Q. Zhu, L.-J. Lou, and N.-Z. Xiao, "Deep learning-based multiscale pyramid sieve and analysis module in image segmentation," *IEEE Access*, vol. 11, pp. 22307–22319, 2023.
[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[49] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021.
[50] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2-FPN for semantic segmentation of fine-resolution remotely sensed images," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1131–1155, Feb. 2022.
[51] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.
[52] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 5607713.
[53] 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, pp. 1–5, 2022.
[54] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.
[55] L. Bai, J. Yang, C. Tian, Y. Sun, M. Mao, Y. Xu, and W. Xu, "DCANet: Differential convolution attention network for RGB-D semantic segmentation," 2022, *arXiv:2210.06747*.
[56] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, p. 3065, Aug. 2021.
[57] J. Wang, Z. Feng, Y. Jiang, S. Yang, and H. Meng, "Orientation attention network for semantic segmentation of remote sensing images," *Knowl.-Based Syst.*, vol. 267, May 2023, Art. no. 110415.
[58] R. Lin, Y. Zhang, X. Zhu, and X. Chen, "Local–global feature capture and boundary information refinement Swin transformer segmentor for remote sensing images," *IEEE Access*, vol. 12, pp. 6088–6099, 2024.
[59] Q. Song, J. Li, C. Li, H. Guo, and R. Huang, "Fully attentional network for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2280–2288.
[60] Y. Quan, D. Zhang, L. Zhang, and J. Tang, "Centralized feature pyramid for object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 4341–4354, 2023.
[61] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[62] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819.
[63] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-Mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.
[64] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

YUANYUAN DANG received the master's degree in engineering from Changchun University of Technology, in 2006. She is currently working as an Associate Professor with the School of Computer Science and Engineering, Changchun University of Technology. Her research interests include artificial intelligence, deep learning, and remote sensing image processing.

YU GAO received the bachelor's degree in computer science and technology from Changchun University of Technology, in 2022, where she is currently pursuing the master's degree in computer technology with the School of Computer Science and Engineering. Her current research interests include remote sensing image semantic segmentation and image processing.

BING LIU received the master's degree in computer application technology from Changchun University of Technology, in 2007. He is currently pursuing the Ph.D. degree in computer application technology with Jilin University. He is currently working as an Associate Professor and the Master's Supervisor with Changchun University of Technology. His research interests include artificial intelligence, computer vision, information security, and deep learning.