Semantic Segmentation of Remote Sensing Images Based on Filtered Hybrid Attention Mechanisms

Sunan Ge, Daihua Liu, Xin Shi, Xueqing Zhao, Xinying Wang, Jianchao Fan,

***Abstract*—Semantic segmentation plays a crucial role in understanding and interpreting of high spatiotemporal resolution remote sensing, which are widely used in many fields such as agriculture, meteorology, and military. With the continuous advancement of convolutional neural networks (CNNs), compared to shallow network learning, the performance of pixel-level classification accuracy as well as multi-scale feature extraction has been greatly improved with deep learning (DL). However, in high spatiotemporal resolution and complex scenarios, the semantic segmentation of remote sensing still faces many problems, such as high computational complexity, easy loss of detail information, and low segmentation accuracy. To address these issues and enhance semantic segmentation accuracy in complex scenarios, in this work, a filtered hybrid attention mechanism network is proposed to provide a robust and effective multi-scale and multi-granularity backbone system for semantic segmentation within the encoder-decoder framework. Firstly, the lightweight network, MobileNetV2 is selected, which can efficiently extract the high-level and low-level features of remote sensing. Meanwhile, a multi-scale filtered hybrid module is constructed for the spatio-temporal spectral multilayer features extraction, which can accurately capture the spatial dependence and spatio-temporal globality of high-level features with the attention mechanism, so as to obtain a fine depiction of the remote sensing features. Furthermore, the accuracy of coarse-grained features is improved by capturing the contextual information of low-level features with the help of multi-head attention mechanism. Finally, the results demonstrate that our method outperforms other methods on the LoveDA urban dataset, the ISPRS Vaihingen dataset, and the GaoFen-3 (GF-3) satellite synthetic aperture radar (SAR) dataset, indicating its effectiveness in performing semantic segmentation for remote sensing in complex scenes.**

***Index Terms*—remote sensing, semantic segmentation, filtered hybrid attention mechanism, multi-scale feature.**

I. Introduction

Manuscript received July 24, 2024; revised November 28, 2024. This work was supported by National Social Science Foundation of China Art Project(No.23EH232); Key Research and Development Program of Shaanxi Province in 2023 (No.2023-YBGY-404, No.2023-ZDLGY-48); Shaanxi Natural Science Youth Foundation Program 12(No.2021JQ-694); State Environmental Protection Key Laboratory of Coastal Ecosystem(202110).

Sunan Ge is a lecturer at Shaanxi Key Laboratory of Clothing Intelligence and State-Province Joint Engineering and Research Center of Advanced Networking and Intelligent Information Services at the Computer Science School, University of Xi'an Polytechnic, Shaanxi, 710048, P. R. China. (corresponding author, e-mail: gesunan@xpu.edu.cn).

Daihua Liu is a postgraduate student at the Computer Science School, University of Xi'an Polytechnic, Shaanxi, 710048, P. R. China. (e-mail: 220721102@stu.xpu.edu.cn).

Xin Shi is an engineer at the Computer Science School, University of Xi'an Polytechnic, Shaanxi, 710048, P. R. China. (e-mail: shixin@xpu.edu.cn).

Xueqing Zhao is an associate professor at the Computer Science School, University of Xi'an Polytechnic, Shaanxi, 710048, P. R. China. (e-mail: zhaoxueqing@xpu.edu.cn).

Xinying Wang is a director at Big Data Application Research Section of CEPRI, China Electric Power Research Institute, Beijing, 100192, P. R. China. (e-mail: wangxinying@epri.sgcc.com.cn).

Jianchao Fan is a professor at the Control Science and Engineering School, Dalian University of Technology, Liaoning, 116024, P. R. China. (e-mail: fjchao@dlut.edu.cn).

REMOTE sensing [1] is a type of Earth observation data characterized by wide chronological coverage, rich spectral information, and variable target structure, which provides a more comprehensive characterization of the surface for many tasks in earth science research, including change detection, land cover mapping, object extraction, and other tasks. With the continuous development of high spatiotemporal resolution imaging technology, the range of remote sensing observations is further expanded to express geospatial information in a finer way. These geospatial information plays an important role in urban and rural planning and construction, transportation analysis and prediction, and military target identification [2]. In particular, by understanding the feature information in remote sensing, classifying the pixels in remote sensing according to the feature categories represented, and providing accurate and detailed geospatial information for the application fields of remote sensing is an important goal of semantic segmentation. Due to the high resolution and rich spectral information of remote sensing, similar appearance features exist between features such as buildings, vegetation and water bodies, light changes, shadows and occlusions. Therefore, how to extract effective information and distinguish similar appearance features from massive remote sensing observation data is a key task in semantic segmentation of high spatiotemporal resolution remote sensing.

Traditional semantic segmentation methods mainly segment an image into several disjoint regions by extracting the low-level features of the image (e.g., grayscale, color, spatial texture, and geometric features used for remote sensing). According to the consistency or similarity exhibited by the low-level features within the same region and the significant differences between different regions, a threshold method is adopted to divide the image into different regions. In addition, for improving the accuracy of regional refinement refine the regions, watershed segmentation algorithm based on the mathematical morphology of topological theory [3], normalized segmentation algorithm based on the global information of the image [4], and the ideological fuzzy C-means based on clustering algorithms [5] have been successively proposed. These segmentation methods, which rely on low-level features, primarily group pixels into meaningful objects based on predefined parameters and achieve the semantic segmentation process by computing additional features such as texture, context, and shape relevance as a set of classification features [6]. During the classification process, traditional low-level feature extraction methods only consider the value of each pixel while neglecting the connections between the current pixel and its surrounding pixels, which cannot meet the requirements of remote sensing interpretation. As high spatiotemporal resolution remote sensing contain increasingly richer details and more complex backgrounds,

it becomes challenging to capture the high-level semantic information of remote sensing in complex environments using low-level feature extraction methods. Therefore, to a certain extent, the limitations of remote sensing semantic segmentation methods based on low-level features are low classification accuracy and poor adaptability to complex samples.

With the impressive performance of image processing techniques based on deep convolutional neural networks(CNNs) in semantic segmentation tasks, they have been widely applied in the field of remote sensing semantic segmentation [7]. Taking advantage of powerful feature representation and data fitting ability of deep CNNs to access high-level image semantic information [8], Long et al. proposed a fully convolutional network (FCN) [9], which is an end-to-end training network that replaces the fully-connected layer in traditional CNNs with a convolutional layer, and directly performs pixel-by-pixel predictions on inputs of any size through upsampling layers and the output layer within the network. However, due to the presence of pooling layers, multiple convolution, and successive pooling operations, the network continuously expands the receptive field and aggregates information [10], resulting in a lack of spatial consistency during downsampling, which reduces the spatial resolution of the image and ultimately loses global contextual information.

To further acquire contextual information, Ronnerberge et al. proposed U-net network [11], which designs two paths to learn contextual and spatial information by introducing an encoder-decoder paradigm to obtain feature maps that are combined within different dimensions to form denser features. Moreover, by utilizing skip-connection modules, the network is allowed to transmit information between different layers, thereby addressing the issue of missing contextual information and enabling it to acquire both high-level semantic information and low-level detailed information simultaneously. However, due to the complex skip-connection mechanism of the deep network structure [12], [13], [14], the training speed of U-net is relatively slow, which may lead to overfitting of the network, thus reducing the generalization ability. In addition, when facing the complex ground information contained in remote sensing, many objects have similar appearance and the task is easily confused. Then, for improving the accuracy of complex information segmentation and fully utilizing the spatial context information, Zhao et al. proposed a pyramid scene parsing network (PSPNet) [15], which employs a pyramid pooling module (PPM) module to add local and global information to the feature map, enabling the model to consider more global contextual information, improving the performance of capturing global information, and performing multi-scale feature fusion. The auxiliary loss function is also added to make the convergence speed increase when training the network. The results demonstrate that the model effectively improves the accuracy of semantic segmentation.

For enhancing the ability to capture image details and multi-scale contextual information, Chen et al. proposed the DeepLabV3 plus [16], which adopts an encoder-decoder structure based on the atrous spatial pyramid pooling (ASPP) module to improve the detection speed of the network. By using different dilation rates, it extracts features from various receptive fields, expanding the receptive field to capture rich information for precise segmentation results. The PPM and ASPP module are capable of considering more contextual information, enhancing multi-scale feature representation, and enabling multi-scale features to cover a broader and denser range of scales, thereby effectively improving the accuracy of semantic segmentation [17], [18]. Unfortunately, both DeepLabV3 plus and the PSPNet encounter challenges in fully harnessing high-level multi-scale feature information during image feature extraction. This limitation results in the loss of vital details, ultimately impacting the precision of segmentation.

The residual ASPP with attention framework network (RAANet) [19] proposed by Liu et al. introduces residual blocks in the ASPP module and reconstructs the null convolutional units using extended attention convolutional units to capture important semantic information at multiple scales. This design allows the network to capture features at different scales more efficiently, thus improving segmentation performance. Also, by introducing residual units, the complexity of the network is reduced, making the training and inference process more efficient. In this way, RAANet achieves better results in semantic segmentation tasks, enabling the network to understand and interpret semantic information in high-resolution remote sensing images accurately. However, its original primitive feature extraction network Xception [20] has a large number of parameters, resulting in a relatively complex model, which is not favorable for deployment in scenarios with limited computational resources, and thus is not effective in the face of small data samples. Furthermore, the attention mechanism in the original model may still have some limitations when dealing with complex scenarios, making it difficult to adequately capture semantic information of various scales and complexities.

Based on these limitations, proposed with a remote sensing images semantic segmentation model based on a filtered hybrid attention mechanism

1. Firstly, the lightweight MobileNetV2 was chosen to extract both low-level and high-level features, and further optimized on this basis to address the issues of spatial detail loss and inadequate feature extraction.

2. Secondly, a filtered hybrid attention mechanism module was designed for use in the high-level feature part of the backbone network. In order to enhance the model's ability to pay attention to features at different scales and to obtain global contextual information about the entire image, thereby improving the model's understanding of object boundaries and details.

3. Finally, a multi-head attention mechanism module was employed for the low-level feature part to enhance the accuracy of semantic segmentation.

II. Methods

A. Filtered Hybrid Attention Mechanism Module

In order to solve the limited expression capability of a single attention mechanism in capturing global information features and contextual information, the image detail expression is enhanced by introducing the 2D Fourier Transform [21], while a filtered hybrid attention mechanism is designed by fusing the Convolutional Block Attention Module (CBAM)

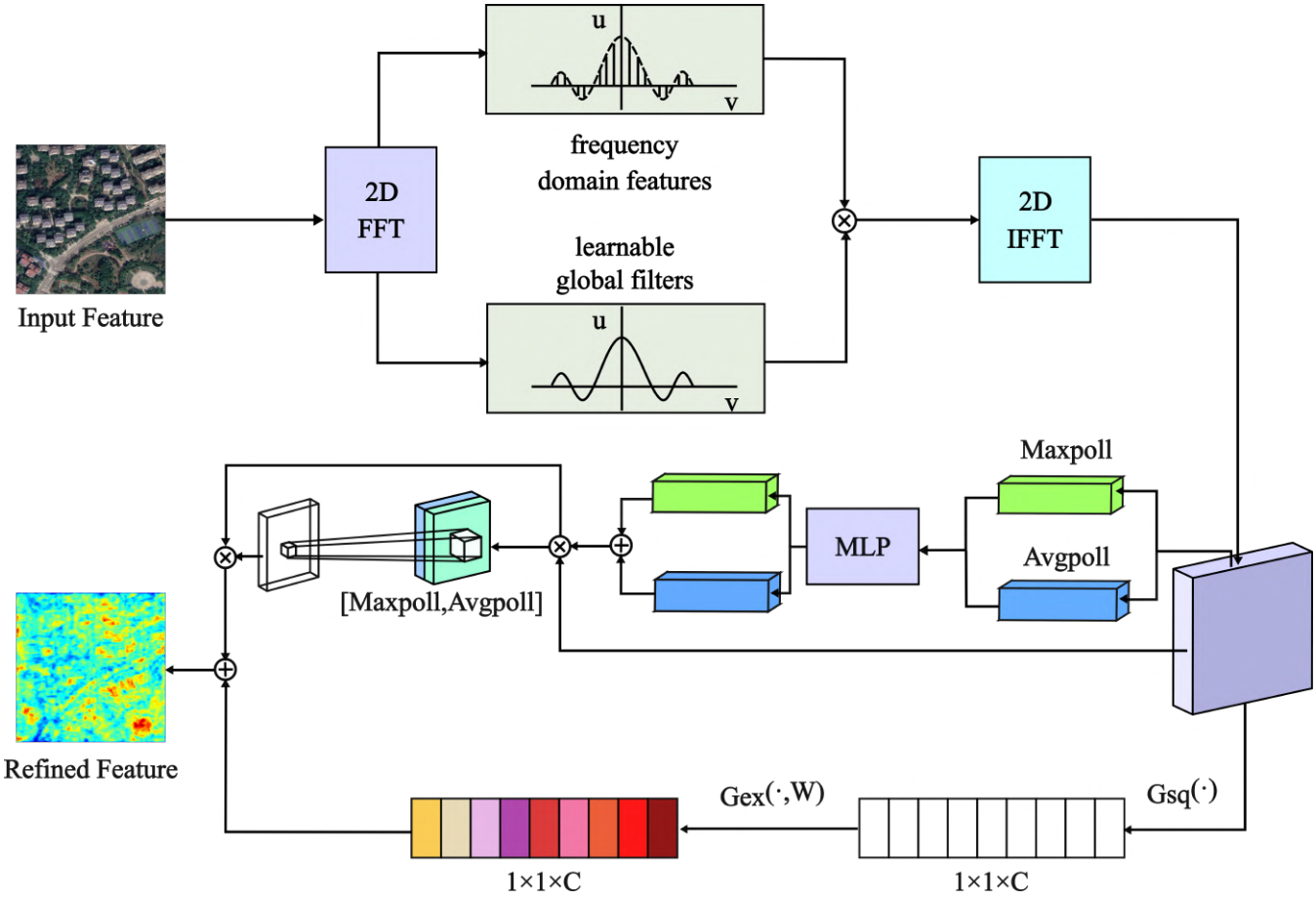

Fig. 1. Module diagram of the filtered hybrid attention mechanism.

[22] and the SE module [23], which solves the problem of the spatial features ignored by the SE module that primarily focuses on channel information, respectively. Furthermore, since CBAM lacks sufficient access to channel information, a filtered hybrid attention mechanism is able to better capture global contextual information and enhance detailed feature extraction capabilities, thereby improving the accuracy and generalization of the model. Its structure is shown in Fig.1.

Typically, 2D Fourier transforms include both continuous Fourier transforms and discrete Fourier transforms, where the continuous Fourier transform is performed as follows:

$$F(u,v) = \int \int_{-\infty}^{\infty} f(x_n c, x_m) \cdot e^{-2\pi i(un+vm)} dx_m dx_n, m \tag{1}$$

in which u, v denote the frequency coordinates in the frequency domain, respectively. N, M and C are the height, width and number of channels of the feature maps of the input image X. n and m denote the new pixel point position and c denotes the channel of x, respectively. The discretized 2D Fourier transform is defined by:

$$F(u,v) = \sum_{n=1}^{N} \sum_{m=1}^{M} X_{n,m} \cdot e^{-j2\pi(\frac{un}{N}+\frac{vm}{M})}, \tag{2}$$

frequency domain obtained after discretization:

$$\bar{X} = F(u,v) \in C^{C\times N\times M}, \tag{3}$$

where $\bar{X}$ is a complex tensor representing the spectrum of X. A learnable filter $K \in C^{C\times N\times \mathrm{M}}$ is then multiplied by $\bar{X}$ to modulate the spectrum:

$$\tilde{X} = K \odot \bar{X}, \tag{4}$$

where $\odot$ is the multiplication of elements, the filter K has the same dimension as $\bar{X}$, which is referred to as the global filter, and $\tilde{X}$ represents filters in the frequency domain.

Finally, we use the inverse Fourier transform to modulate the spectral $\tilde{X}$ transformed back to the spatial domain:

$$G = g(i,j) = \int \int_{-\infty}^{\infty} \tilde{X} \cdot e^{2\pi i(\mathrm{un}+\mathrm{vm})} dudv, \tag{5}$$

where $g(i,j)$ denotes the position of the feature image G. The spatial dependence of the image and the positional details of the feature G are obtained after global filtering.

In order to further extract the channel and spatial key feature information of the image features and remove the redundant information, a hybrid attention mechanism is employed here. Firstly, the channel features are extracted by the channel attention module of the CBAM, and the input feature map G undergoes global average pooling and global max pooling processes to generate two $1 \times 1 \times C$ feature maps. Secondly, through the multilayer perceptron (MLP) for feature summation, to obtain the corresponding $M_{\mathrm{c}}(G)$, and then multiply with the input feature map G. Finally, obtain the input features G' of the spatial attention module, which is mathematically formulated as follows:

$$M_{\mathrm{c}}(G) = \sigma(\mathrm{MLP}(\mathrm{Avg}(G)) + \mathrm{MLP}(\mathrm{Max}(G))), \tag{6}$$

$$G' = M_{\mathrm{c}}(G) * G, \tag{7}$$

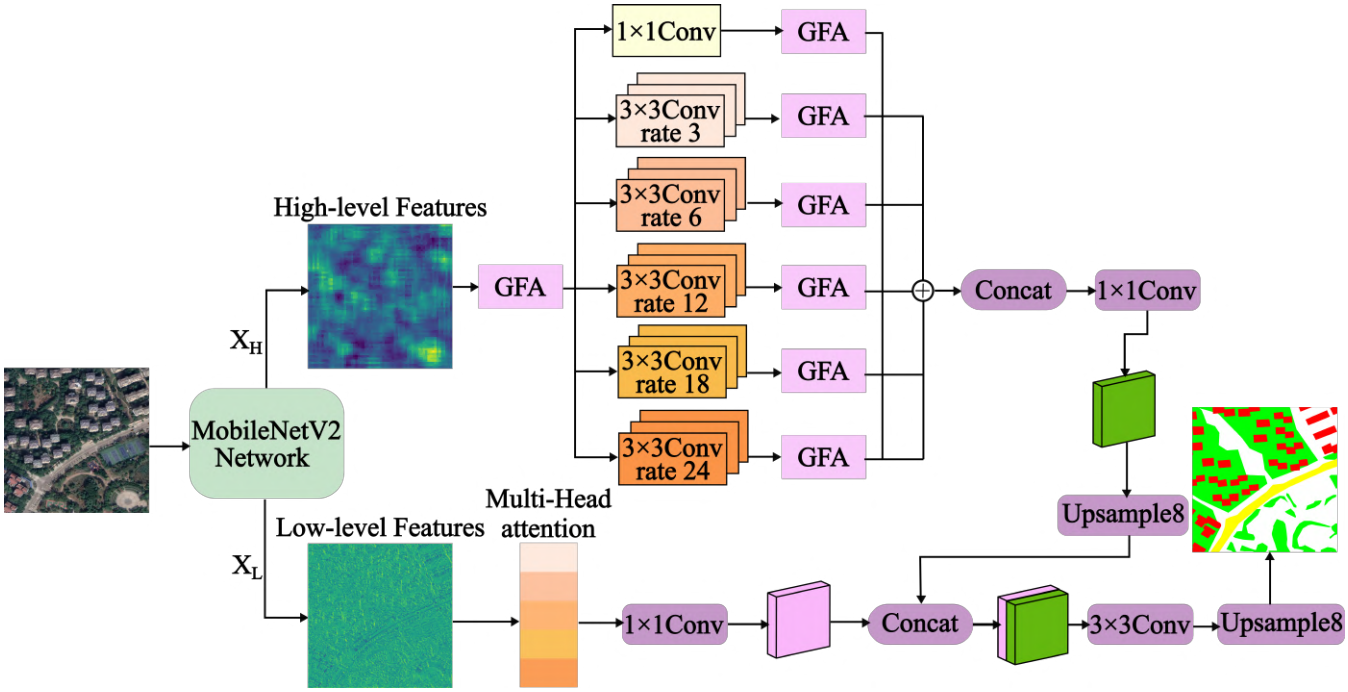

Fig. 2. The Framework diagram of semantic segmentation network based on filtered hybrid attention mechanism.

where σ is the sigmoid operation.

Further, the channel features obtained above are subjected to global max pooling and global average pooling to obtain two $N \times M \times 1$ feature maps. The corresponding channel splicing operation is performed on the two feature maps to obtain the $N \times M \times 2$ feature maps, which are then processed by 7×7 convolutional downscaling and activation function to generate the $N \times M \times 1$ spatial features $M_s(G')$. Similar to the channel feature extraction process, the spatial features $M_s(G')$ are multiplied with the input feature maps G' to generate the final spatial feature maps G'', as follows:

$$M_S(G') = \sigma(f^{7*7}([\mathrm{Avg}(G'); \mathrm{Max}(G')])), \tag{8}$$

$$G'' = M_S(G') * G', \tag{9}$$

where 7*7 in f^{7*7} denotes the size of the convolution kernel, and the feature map G' output from the channel attention module is used as the input feature map for this module.

In order to further extract the interdependence information between channels, the feature map G of $N \times M \times C$, which contains global information, is compressed into a feature vector Z of $1 \times 1 \times C$, and the feature data Z_c with contextual information is obtained, as shown in Equation 10:

$$Z_c = G_{sq}(G) = \frac{1}{N * M} \sum_{i=1}^{N} \sum_{j=1}^{M} g(i,j), \tag{10}$$

where Z_c is the compression feature of channel c.

For determining the dependency between channels, the importance between different channels is obtained by learning two fully connected layers to calculate the weight of each channel based on compression, i.e.,

$$f_c = \delta(\mathrm{F}_1(Z_C)), \tag{11}$$

where f_c is the feature after nonlinear mapping, F_1 denotes the fully connected layer, and δ is the ReLU activation function. Then f_c is subjected to a second full-connectivity learning to obtain the excitation weights for each channel as follows:

$$S_c = \sigma(\mathrm{F}_2(f_c)), \tag{12}$$

where S_c is the excitation weight of channel c, F_2 denotes the fully connected layer, and σ is the Sigmoid activation function. and the input feature map G is weighted by

$$Y = S_c * G. \tag{13}$$

The final hybrid feature image M obtained by summing the channel-space feature map G'' and the multichannel characteristic information Y, respectively, is as follows:

$$M = Y(i,j) + G''(i,j), \tag{14}$$

B. Semantic Segmentation Network Based on Filtered Hybrid Attention Mechanisms

A semantic segmentation network has been proposed based on the filtered hybrid attention mechanism module, as illustrated in Fig.2.

The input image on this network structure first passes through the MobileNetV2 network to obtain low-level and high-level features. For enhancing the model's ability to understand the global semantic information of the image, a Filtered Hybrid Attention Mechanism (FHAM) is introduced to enhance the model's ability to process the global information of the whole image. The low-level feature part of MobileNetV2 is mainly responsible for capturing the detail and texture information of the image but has some deficiencies in the global semantic understanding the introduction of the multi-head Attention mechanism allows the model to focus more on the important semantic regions in the image at the low-level feature level, which enhances the delivery of semantic information. Additionally, the incorporation of the multi-head attention mechanism enables the model to better learn feature representations with semantic properties.

By weighting the attention to different spatial regions of low-level features, the model is able to capture semantic information at different scales in the image more accurately, thus improving the semanticity of the features. This improvement allows the model to perceive a wider range of contextual information, which helps to improve the model's ability to understand the subtle structures and boundaries in the image, resulting in a significant improvement in the accuracy of semantic segmentation.

1) MobileNetV2 network

MobileNetV2 [24] is improved by adding inverted residuals with a linear bottleneck module to MobileNetV1. The MobileNet architecture is based on depth-separable convolution, where standard 2D convolution produces an output channel by processing all input channels directly in the depth dimension (channels).

Depth-separable convolution, on the other hand, divides the input image and filters into separate channels, and then convolves each input channel with the corresponding filter channel. After generating the filtered output channels, these output channels are then stacked back. In separated deep convolution, the stacked output channels are filtered using a 1×1 convolution, called point-by-point convolution, which combines the stacked output channels into a single channel. Deep separable convolution produces the same output as standard convolution, but is more efficient due to the reduction in the number of parameters involved. MobileNetV2 inserts 19 inverted residual bottleneck layers after the first convolutional layer with 32 filters, and then finishes with a point-by-point convolution that produces a size of 7 × 7 × 1280 pixels. The residual block connects the beginning and the end of the convolution block via jump connections, with the aim of passing information to deeper layers of the network. In a standard residual block, the beginning and end of the convolution block typically have more channels than the middle layer. In the inverted residual block used in MobileNetV2, the connected layers have fewer channels than the intermtediate layers, yielding much fewer parameters than the standard residual block.

TABLE I
MOBILENETV2 NETWORK.

Input	Operator	t	c	n	s
224^2×3	conv2d	-	32	1	2
112^2×32	bottleneck	1	16	1	1
112^2×16	bottleneck	6	24	2	2
56^2×24	bottleneck	6	32	3	2
28^2×32	bottleneck	6	64	4	2
14^2×64	bottleneck	6	96	3	1
14^2×96	bottleneck	6	160	3	2
7^2×160	bottleneck	6	320	1	1
7^2×320	conv2d 1x1	-	1280	1	1
7^2×1280	avgpool 7x7	-	-	1	-
1×1×1280	conv2d 1x1	-	k	-	

Each line describes a sequence of 1 or more identical layers, repeated n times. All layers in the same sequence have the same number c of output channels. The first layer of each sequence has a stride s and all others use stride 1. All spatial convolutions use 3×3 kernels. The expansion factor t is always applied to the input size.

The paper is based on the MobileNetV2 network architecture of Table I for low-level feature and high-level feature extraction of remote sensing images.

2) Multi-head Attention Mechanism

Multi-head Attention Mechanism (MAM) [25] is a network structure based on self-attention [26] for enhancing the degree of attention of a neural network to different parts or different levels of features, allowing the network to learn multiple attentional weights at the same time and apply these weights to different feature representations. Its mathematical structure is as follows, first for the self-attention mechanism is defined as shown in equation (15):

$$\mathrm{Att}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{15}$$

where Q, K, V are multidimensional vectors (each of its dimensions is q_i, k_i, v_i respectively), which are computed by multiple input nodes through W_q, W_k, W_v transformation matrices respectively. The process of matching q and k can be regarded as calculating the correlation between the two, and the larger the correlation, the larger the weight of the corresponding v. Q dot product K^T for q and k matching, QK^T is divided by $\sqrt{d_k}$ (d represents the length of vector k) to scale the dot product result, and then processed by Softmax function to get the weight for v, and then multiplied by vector V for weighting to get the final result.

In order to correspond to the semantic segmentation network structure, the self-attention mechanism designed in this paper is:

$$\mathrm{Att}(X_{L1,}X_{L2},X_{L3})=\mathrm{softmax}\left(\frac{X_{L1}X_{L2}^T}{\sqrt{d_k}}\right)X_{L3}, \tag{16}$$

where $X_L=\{x_{n,m,c}\}$, based on the self-attention basis, then constructs the multi-head attention mechanism as follows:

$$\mathrm{Multi}(X_{L1,}X_{L2},X_{L3})=\mathrm{Concat}(h_1,h_2,...,h_i)W^O, \tag{17}$$

$$h_i=\mathrm{Att}(X_{L1}{W_i}^Q,X_{L2}{W_i}^K,X_{L3}{W_i}^V), \tag{18}$$

where i is the number of head, W^O is the learnable parameter.

III. EXPERIMENTATION AND RESULTS

A. Datasets

In order to fully evaluate the effectiveness of the proposed method, three datasets are used to validate the effectiveness of the proposed method in semantic segmentation of multi-categorized complex scenes and semantic segmentation of a small number of datasets. The urban area portion of the LoveDA dataset and the ISPRS Vaihingen dataset are validated for multi-categorized complex scenarios, and the data from raft farming areas are used to validate the under-categorized small number of sample scenarios.

1) LoveDA dataset

The Land-cOVEr Domain Adaptive (LoveDA) semantic segmentation dataset [27] is used to validate the semantic segmentation of the proposed method in complex scenarios, which contains 5987 remote sensing images with a spatial resolution of 3m (high), and the resolution of each image is 1024 × 1024. The dataset includes remote sensing images from different urban and rural areas in Nanjing, Changzhou and Wuhan. It encompasses six land use categories, namely building, road, water, barren land, forest, and agriculture.

Given its intricate background, multi-scale objects, and variable class distribution, this dataset presents a formidable challenge for semantic segmentation tasks. In this study, we utilise the urban subset of this dataset to assess the model's performance.

2) ISPRS Vaihingen dataset

The ISPRS Vaihingen dataset [28] offers a cutting-edge compilation of aerial imagery, optimized for urban classification and 3D building reconstruction initiatives. The dataset features a digital surface model (DSM) derived from high-resolution orthophotos and dense image matching techniques, which encompass a variety of urban scenes. The dataset primarily represents a small village with numerous individual structures and low-rise buildings. The dataset comprises 33 remote sensing images of varying dimensions, each extracted from a larger top-level orthophoto. The compilation process ensures the exclusion of areas without data. The remote sensing images are presented in the 8-bit TIFF file format, comprising three bands: near-infrared, red, and green. Due to the considerable dimensions of the original images, which would render direct use impractical, they have been processed to conform to a standardized dataset format. Specifically, the original images were cropped to produce 3,300 images, each 512×512 pixels in size. The dataset categorizes these images into six primary land cover classes: impervious surfaces, buildings, low vegetation, trees, cars, and background.

3) Remote sensing image dataset of raft farming area

This paper presents a limited number of samples from semantic segmentation experiments, primarily utilizing Gaofen-3 (GF-3) satellite synthetic aperture radar (SAR) remote sensing images of a specific sea area within the raft aquaculture region. The remote sensing images exhibit a spectral composition of R, G, and B, with a spatial resolution of less than one meter. The dataset comprises four high-resolution SAR remote sensing images, each of which has a different size. The largest image is 10724 × 8040, while the smallest is 10525 × 8048. As the images are too large to be used directly, they must be cropped to the standard dataset format. To facilitate the image processing, the segmented network model can be trained and features extracted from the image with greater accuracy. Additionally, the position of the ocean rafts can be more effectively displayed. The original image was cropped to obtain an image of size 416×416.

B. Evaluation metrics

In order to effectively evaluate the performance of the proposed method, metrics such as mean Intersection over Union (mIoU), mean Recall (mRecall), mean Precision (mPrecision), and F1-Score are used to assess the overall performance of the different models.

IoU is the ratio of the intersection and concatenation of the predicted results with the true values. mIoU is a standardized evaluation of the average IoU of all types with the following formula:

$$IoU = \frac{TP}{TP + FN + FP}, \tag{19}$$

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP + FN + FP}, \tag{20}$$

where True Positive (TP): indicates that the model correctly predicts the number of positive class samples as positive class samples. False Positive (FP): indicates that the model incorrectly predicts the number of negative class samples as positive class samples. False Negative (FN): indicates the number of positive class samples that the model incorrectly predicts to be negative class samples.

Recall is used to evaluate the ability of the classifier to find all positive samples, and mRecall is the average recall across all types, calculated as follows.

$$Recall=\frac{TP}{TP + FN}, \tag{21}$$

$$mRecall = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP + FN}. \tag{22}$$

Precision denotes the ability of the classifier to label a sample as positive, while mPrecision is the average precision across all types, calculated as follows:

$$Precision=\frac{TP}{TP + FP}, \tag{23}$$

$$mPrecision = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP + FP}. \tag{24}$$

The F1-Score is defined as the reconciled mean of recall and accuracy; it focuses on precision and recall and provides an overall measure of the performance of a change detection model. The higher the F1-score, the more accurate the performance, which is calculated as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{25}$$

C. Ablation experiment

The ablation experiments on the LoveDA dataset are validated based on the MobileNetV2, MobileNetV2+ FHAM and MobileNetV2+ FHAM +MAM backbone networks, and the performance metrics are shown in Table II.

Through the experimental results, it can be seen that the added FHAM module, compared with the backbone network, improves the mIoU value by 3.83% and the F1 value by 1.92%, which can show that the FHAM is able to extract the detail information of the remote sensing image very well, and furthermore, the added FHAM module and the MAM module improve the mIoU value by 0.96% and the F1 value by 0.52%, which is able to capture the shallow semantic information of the remote sensing images.

Ablation experiments were also performed on the ISPRS Vaihingen dataset and the performance metrics are shown in Table III.

The experimental results show that the addition of the FHAM module increases the mIoU value by 1.01% and the F1 value by 0.48% compared to the backbone network, and the addition of the FHAM module and the MAM module increases the mIoU value by 0.33% and the F1 value by 0.18%.

D. Segmentation Experiments on the LoveDA Dataset

In order to verify the effectiveness of the proposed method in remote sensing semantic segmentation in complex scenes, through the LoveDA dataset, the proposed method in this paper and a variety of mainstream network models conducted

TABLE II
ABLATION EXPERIMENT ON THE LOVEDA DATASET.

methods	mIoU	mPrecision	mRecall	F1-Score
MobileNetV2	59.07	74.63	74.34	74.48
MobileNetV2+FHAM	62.90	74.31	78.62	76.40
MobileNetV2+FHAM+MAM	63.86	75.52	78.38	76.92

TABLE III
ABLATION EXPERIMENT ON THE ISPRS VAIHINGEN DATASET.

methods	mIoU	mPrecision	mRecall	F1-Score
MobileNetV2	86.03	91.63	93.35	92.48
MobileNetV2+FHAM	87.04	92.36	93.70	93.02
MobileNetV2+FHAM+MAM	87.37	92.75	93.66	93.20

TABLE IV
METRICS VALUES FOR DIFFERENT SEMANTIC SEGMENTATION MODELS ON THE LOVEDA DATASET.

methods	mIoU	mPrecision	mRecall	F1-Score
PSPNet	52.92	64.98	72.47	68.52
Unet	59.87	71.12	77.45	74.15
DeepLabV3 plus	52.70	69.39	67.39	68.37
RAANET	59.07	74.63	74.34	74.48
Ours	63.86	75.52	78.38	76.92

a comprehensive comparison experiment, and obtained the relevant image segmentation effect as shown in Fig.3, and the performance indexes as shown in Table IV.

Fig.3 shows the results of image segmentation, it can be seen from Fig.4 (a) that PSPNet, Unet, DeepLabV3 plus and RAANET are good for semantic segmentation of most of the buildings in the remote sensing map, however, there are some details that these networks cannot recognize effectively, for example, the house in the lower right corner, which is effectively recognized by the proposed method in this paper.

Fig.3 (b) in the whole scene boundaries is relatively fuzzy, DeepLabV3 plus, RAANET on the forest and wasteland detail portrayal is not enough, resulting in the results compared with other networks is poor, and PSPNet, Unet in the segmentation of buildings above the details of the processing is not good, this paper's proposed method of the details of the two cases segmentation effect is relatively better.

Fig.3 (c) shows that PSPNet, Unet, DeepLabV3 plus and RAANET are able to recognize the corresponding wasteland and buildings in the figure, but the mis-segmentation phenomenon is generated for the parts that do not belong to the labels, and the method proposed in this paper generates less mis-segmentation.

Fig.3 (d) shows that PSPNet, Unet, DeepLabV3 plus and RAANET are able to recognize the red building part of the map, but are not effective in the face of the yellow road area which has a small difference in boundaries, and the method proposed in this paper is able to recognize the road part of the map effectively.

Table IV shows that the proposed method obtained 10.94% higher mIoU value and 8.4% higher F1 value compared to PSPNet, 3.99% higher mIoU value and 2.77% higher F1 value than Unet, 11.26% higher mIoU value and 8.55% higher F1 value than DeepLabV3 plus,4.79% higher mIoU value and 2.44% higher F1 value than RAANET. The results show that all the indexes are better than the other compared networks, thus proving that the network works well in multiclassification complex scenarios.

E. segmentation Experiments on the ISPRS Vaihingen dataset

TABLE V
METRICS VALUES FOR DIFFERENT SEMANTIC SEGMENTATION MODELS ON THE ISPRS VAIHINGEN DATASET.

methods	mIoU	mPrecision	mRecall	F1-Score
PSPNet	70.44	79.19	86.73	82.79
Unet	82.73	90.16	90.69	90.42
DeepLabV3 plus	80.60	87.62	90.95	89.25
RAANET	85.04	90.96	92.79	91.86
Ours	87.37	92.75	93.66	93.20

In order to further validate the performance of the method proposed in this paper for remote sensing semantic segmentation in complex scenarios, further comparative validation is carried out using the ISPRS Vaihingen dataset and on the correlation network model, and the relevant image segmentation effect is obtained as shown in Fig. 4, and the performance indices are shown in Table V.

Fig. 4 depicts the outcomes of image segmentation. As illustrated in Fig. 5 (a) that PSPNet, Unet, DeepLabV3 plus and RAANET are effective for semantic segmentation of the majority of buildings in the remote sensing map. However, the differentiation between low vegetation and trees is not as apparent as desired, and among these methods, PSPNet, Unet, and DeepLabV3 plus are generally effective in segmenting cars. This is effectively recognized by the proposed method presented in this paper.

Fig. 4 (b) demonstrates that PSPNet, Unet, DeepLabV3 plus, and RAANET yield favorable outcomes for the comprehensive segmentation of the image. However, they are less effective in differentiating between low vegetation and trees, where there is minimal variability, and in identifying small targets. In contrast, the proposed method demonstrates proficiency in these tasks.

Fig. 4 (c) illustrates that while PSPNet, Unet, and RAANET are effective at segmenting large buildings, they are less adept at discerning smaller structures. While DeepLabV3 plus is unable to distinguish between buildings and low vegetation, the method proposed in this paper is capable of effectively detecting the corresponding parts.

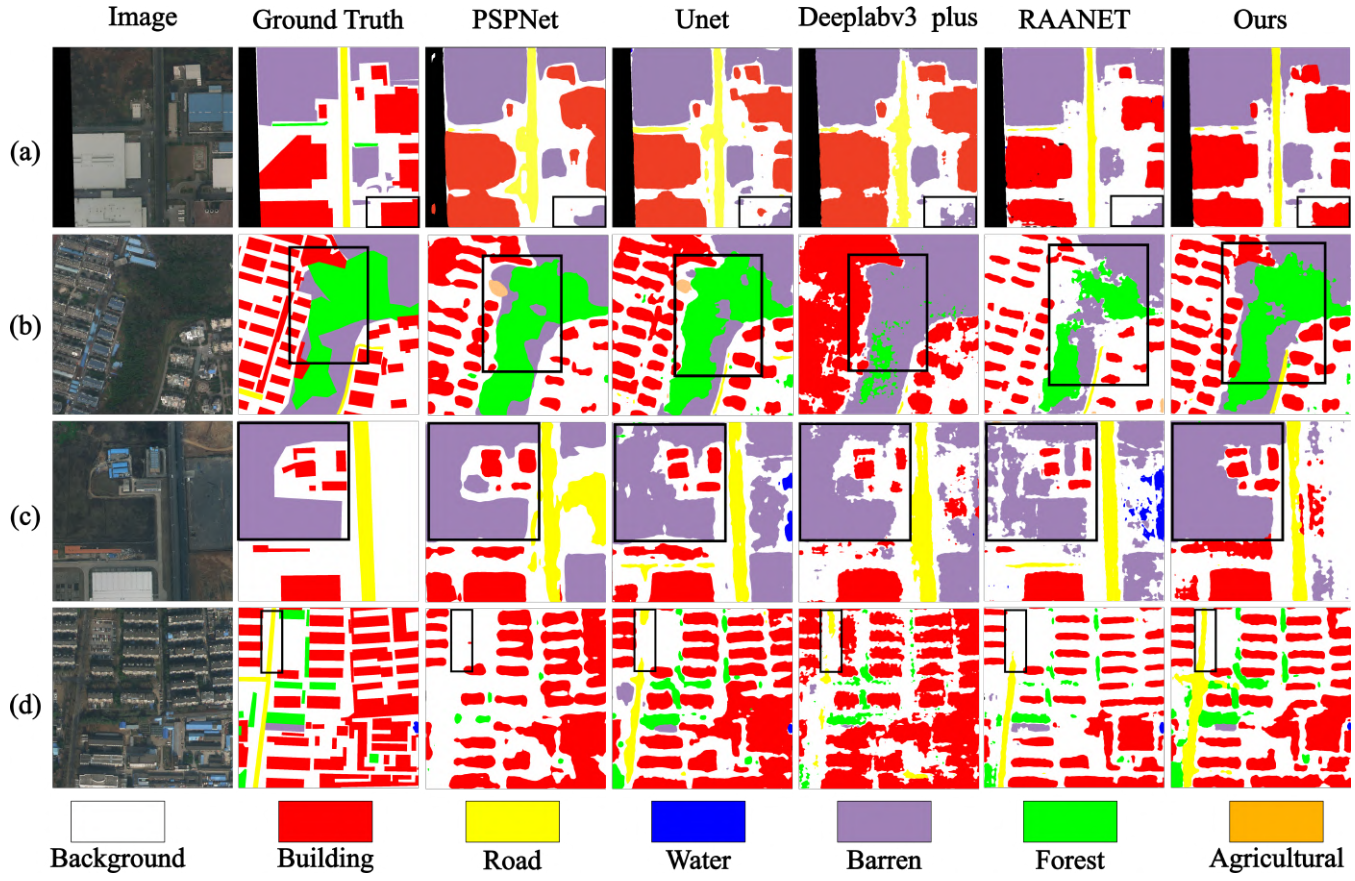

Fig. 3. Comparison of image segmentation results on the LoveDA dataset.

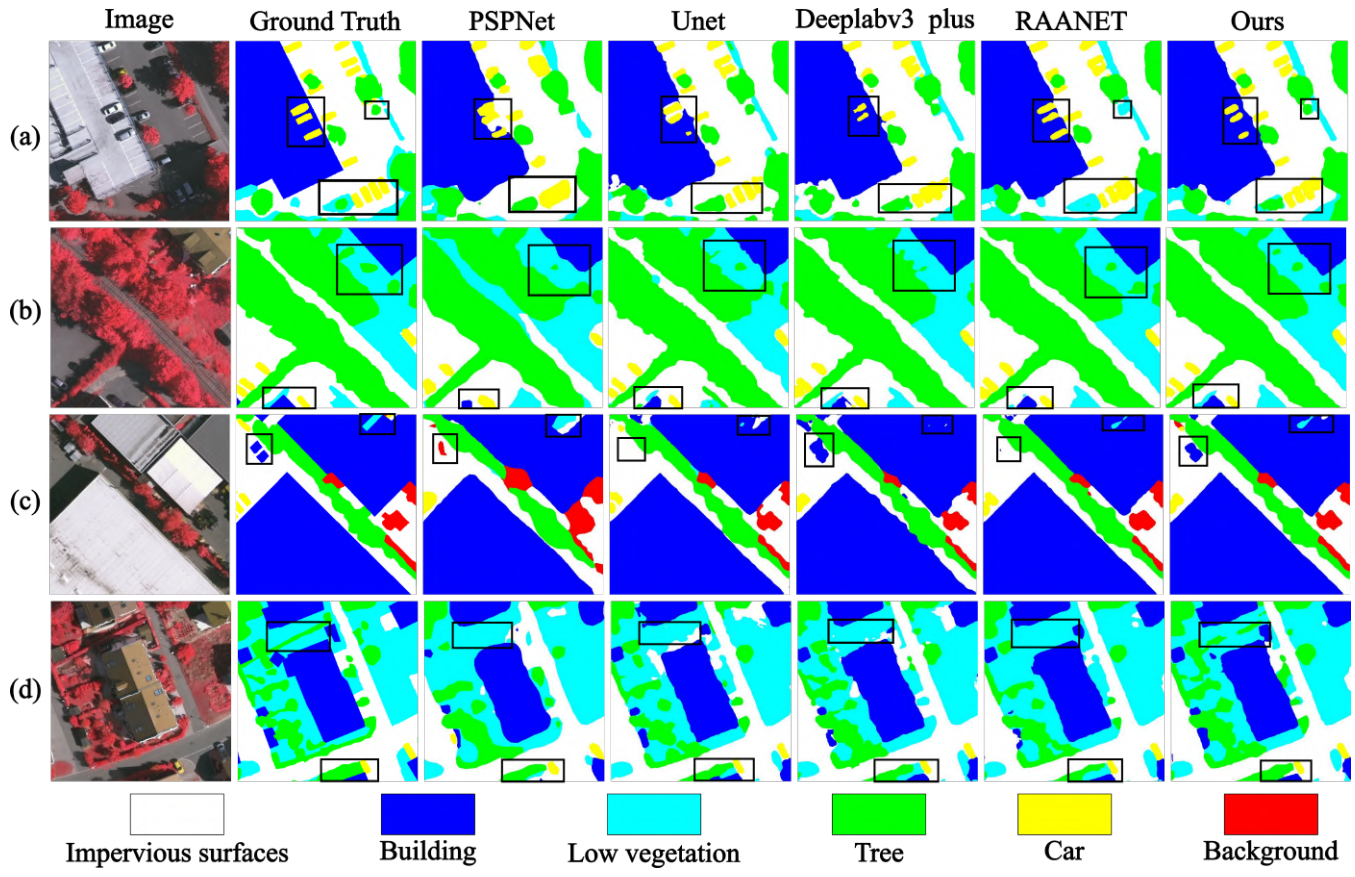

Fig. 4. Comparison of image segmentation results on the ISPRS Vaihingen dataset.

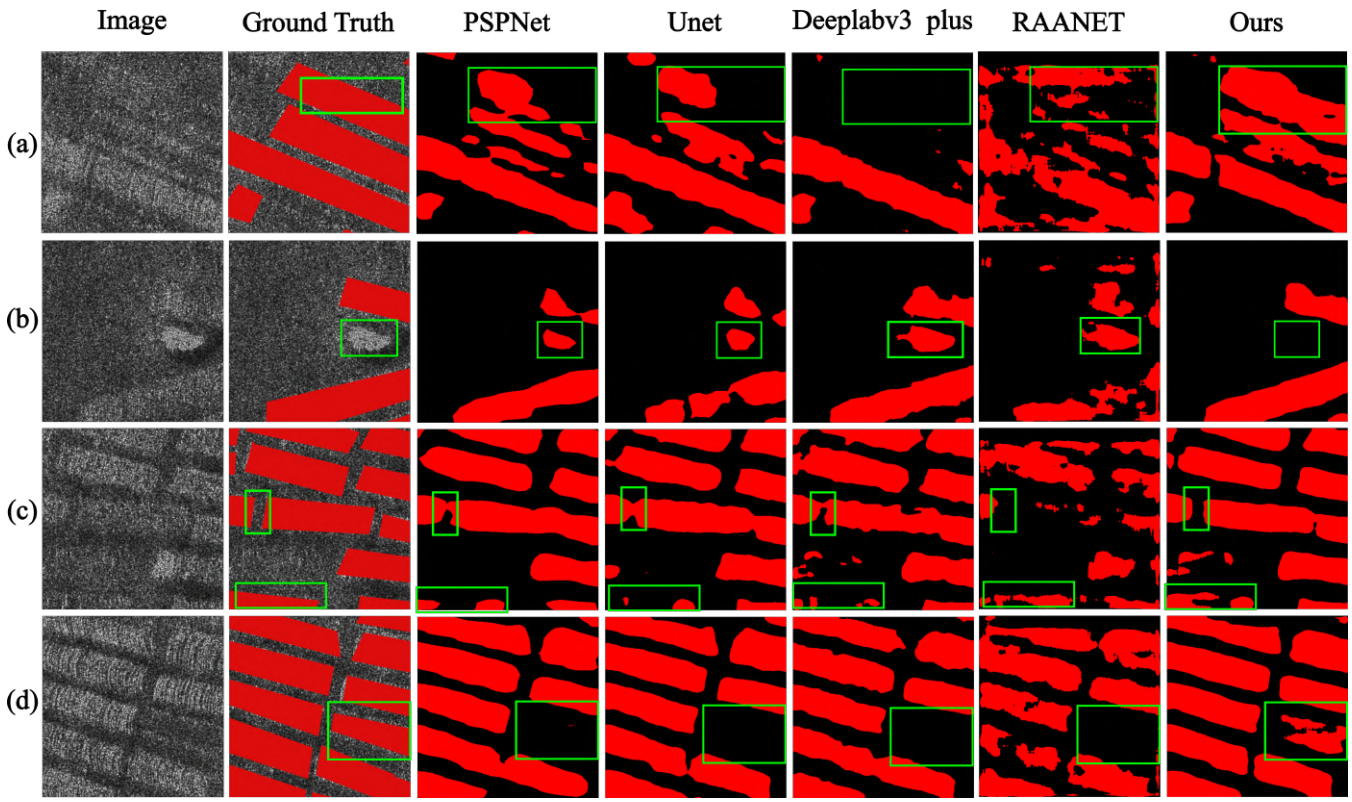

Fig. 5. Comparison of image segmentation results on the remote sensing image dataset of raft culture area.

Fig. 4 (d) demonstrates that while PSPNet, Unet, DeepLabV3 plus, and RAANET are proficient at overall part segmentation, they are not as effective at segmenting parts of complex scenes with low differentiation. In such cases, the method proposed in this paper is more advantageous.

Table V illustrates that the proposed method yielded 16.93% higher mIoU value and 10.41% higher F1 value in comparison to PSPNet, 4.64% higher mIoU value and 2.78% higher F1 value than Unet, 6.77% higher mIoU value and 3.95% higher F1 value than Deeplabv3 plus,2.33% higher mIoU value and 1.34% higher F1 value than RAANET. These results demonstrate that all the indices outperform those of the other comparative networks, thus providing further evidence that the network is more effective for segmentation in multi-categorized complex scenarios.

F. Segmentation Experiments on the Remote sensing image dataset of raft farming area

The remote sensing image dataset in the raft farming area can be taken from a small amount of data with better results, in which the training data is only 159, and the relevant image segmentation effect is obtained as shown in Fig. 5, and the performance indexes are shown in Table VI.

As illustrated in Fig.5 (a), networks such as PSPNet, Unet, DeepLabV3 plus, and RAANET are effective in identifying the clearly classified parts. However, they are less effective in discerning regions with insignificant differences and in segmenting results obtained from small-sample data. In contrast, the network proposed in this paper is effective in identifying target regions.

The Fig.5 (b) illustrates that PSPNet, Unet, DeepLabV3 plus, and RAANET are effective in segmenting the corresponding regions. However, mis-segmentation occurs for regions that do not belong to the labels. In contrast, the method proposed in this paper avoids this phenomenon.

TABLE VI
METRICS VALUES FOR DIFFERENT SEMANTIC SEGMENTATION MODELS ON THE REMOTE SENSING IMAGE DATASET OF RAFT CULTURE AREA.

methods	mIoU	mPrecision	mRecall	F1-Score
PSPNet	77.99	86.50	88.47	87.47
Unet	78.85	86.74	88.17	87.44
DeepLabV3 plus	77.54	86.99	87.21	87.09
RAANET	63.28	76.98	76.47	76.72
Ours	79.92	87.61	89.82	88.70

As illustrated in Fig.5 (c), networks such as PSPNet, Unet, and DeepLabV3 plus yield satisfactory results for the classified regions but exhibit deficiencies in handling intricate details. In contrast, the proposed method demonstrates superior performance in terms of detail-oriented segmentation.

As illustrated in Fig. 5(d), while networks such as PSPNet, Unet, DeepLabV3 plus, and RAANET demonstrate proficiency in overall image segmentation, they tend to exhibit limitations in regions where the segmentation is less apparent. In contrast, the proposed method demonstrates the capacity to perform effective segmentation in such challenging regions.

Table VI demonstrates that our network attains 1.93% higher mIoU values and 1.23% higher F1 values than PSPNet, 1.07% higher mIoU values and 1.26% higher F1 values than Unet, and 2.38% higher mIoU values and 1.61% higher F1 values than DeepLabV3 plus, 16.64% higher mIoU values and 11.98% higher F1 values than RAANET, and it is found that RAANET has poor segmentation results with only a small number of samples. 16.64% and 11.98% higher F1 value, respectively. It is evident that RAANET exhibits

suboptimal segmentation outcomes with a limited number of samples. This substantiates the assertion that our network can achieve superior results with a minimal sample size.

IV. Conclusions

This study proposes an enhanced DL network model for the effective exploration of remote sensing image semantic segmentation. A FHAM is introduced based on MobileNetV2, a lightweight backbone network, to enhance the degree of attention to features at different scales. Furthermore, a MAM is employed in the low-level feature part to improve the accuracy of semantic segmentation. The experimental results demonstrate that our model exhibits significant performance improvements, with a 63.86% increase in mIoU, a 75.52% enhancement in mPrecision, a 78.38% rise in mRecall, and a 76.92% boost in F-score. The 1-score on the LoveDA dataset exhibited an improvement of 4.79%, 0.89%, and 4.04% in comparison to the RAANET model, respectively, with a 2.44% improvement. These findings suggest that the proposed model exhibits superior performance in semantic segmentation tasks and demonstrates enhanced capacity to comprehend and interpret semantic information in high-resolution remote sensing images with greater precision.

The proposed model is capable of more effectively capturing the semantic information present in the image through the incorporation of a filtered hybrid and MAM. These attention mechanisms enable the network to focus more on important features and suppress unimportant features, thereby improving the performance and accuracy of the model. Consequently, the results are of great practical significance for improving the accuracy and efficiency of semantic segmentation of remote sensing images. Furthermore, they can be used as an efficient, accurate, and applicable solution for semantic segmentation of remote sensing images in a variety of scenarios.

References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[2] X. Zhang, L. Han, L. Han, and L. Zhu, "How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery?" *Remote Sensing*, vol. 12, no. 3, p. 417, 2020.

[3] S. Gupta, Y. Zhang, X. Hu, P. Prasanna, and C. Chen, "Topology-aware uncertainty for image segmentation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[4] Y. Wang, X. Shen, Y. Yuan, Y. Du, M. Li, S. X. Hu, J. L. Crowley, and D. Vaufreydaz, "Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[5] J. Jiao, X. Wang, T. Wei, and J. Zhang, "An adaptive fuzzy c-mean noise image segmentation algorithm combining local and regional information∗," *IEEE Transactions on Fuzzy Systems*, 2023.

[6] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021.

[7] K. Yuan, C. Zhao, Y. Chen, L. Shen, Q. Tang, and C. Jia, "Mapping the knowledge structure and research evolution of deep learning," *IAENG International Journal of Computer Science*, vol. 51, pp. 1404–1412, 2024.

[8] J. Hu, L. Li, Y. Lin, F. Wu, and J. Zhao, "A comparison and strategy of semantic segmentation on remote sensing images," in *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 1*. Springer, 2020, pp. 21–29.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[10] Y. Li, H. Zhao, X. Qi, Y. Chen, L. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, "Fully convolutional networks for panoptic segmentation with point-based supervision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4552–4568, 2022.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[12] A. Raza, H. Huo, and T. Fang, "Eunet-cd: Efficient unet++ for change detection of very high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[13] G. Petrakis and P. Partsinevelos, "Lunar ground segmentation using a modified u-net neural network," *Machine Vision and Applications*, vol. 35, no. 3, p. 50, 2024.

[14] H. Jiang and J. Zhao, "Multi-lesion segmentation of fundus images using improved unet+," *IAENG International Journal of Computer Science*, vol. 51, pp. 1587–1595, 2024.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[17] Y. Liu, X. Bai, J. Wang, G. Li, J. Li, and Z. Lv, "Image semantic segmentation approach based on deeplabv3 plus network with an attention mechanism," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107260, 2024.

[18] J. Fu, X. Yi, G. Wang, L. Mo, P. Wu, and K. E. Kapula, "Research on ground object classification method of high resolution remote-sensing images based on improved deeplabv3+," *Sensors*, vol. 22, no. 19, p. 7477, 2022.

[19] R. Liu, F. Tao, X. Liu, J. Na, H. Leng, J. Wu, and T. Zhou, "Raanet: A residual aspp with attention framework for semantic segmentation of high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 13, p. 3109, 2022.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[21] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 980–993, 2021.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[25] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "Multi-head attention: Collaborate instead of concatenate," *arXiv preprint arXiv:2006.16362*, 2020.

[26] Z. Fan, G. Hu, X. Sun, G. Wang, J. Dong, and C. Su, "Self-attention neural architecture search for semantic image segmentation," *Knowledge-Based Systems*, vol. 239, p. 107968, 2022.

[27] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.

[28] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The isprs benchmark on urban object classification and 3d building reconstruction," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I-3, pp. 293–298, 2012. [Online]. Available: https://isprs-annals.copernicus.org/articles/I-3/293/2012/