Sgformer: A Local and Global Features Coupling Network for Semantic Segmentation of Land Cover

Liguo Weng, Kai Pang, Min Xia, *Member, IEEE*, Haifeng Lin, Ming Qian, and Changjie Zhu

***Abstract*—With the introduction of Earth observation satellites, the classification technology through high-definition remote sensing images appeared. After decades of evolution, the land cover classification method in high-definition satellite maps has been gradually improved. Recently, high-definition remote sensing maps have been applied to land cover classification. Nowadays, classification methods using high-definition maps have these following problems. First, the traditional land cover classification methods cannot process the rich details in high-definition maps. Second, there are different acquisition conditions in the maps of different regions, which leads to distortion, deformation, and illumination blur of remote sensing images. Third, the existing methods are unable to provide a good generalization performance. To address these issues, a dual-branch parallel network structure is proposed, called Sgformer, to improve the performance of the transformer in the context of high-definition remote sensing maps. The network enhances perceptual learning with convolution operators that extract local features and a self-attention module that captures global representations. Local information and global representations with semantic divergence are fused through a feature coupling module. At last, a decoder is designed to maximize the preservation of local features and global representations and to better recover high-definition feature maps. The results of semantic segmentation experiments show that the methodology in this study has higher accuracy than the other methodologies.**

***Index Terms*—Deep learning, land cover, neural network, remote sensing, semantic segmentation.**

I. Introduction

With the development of satellites, unmanned aerial vehicles, aircraft, and other remote sensing devices, high-resolution maps are generally applied. Land cover classification contributes to the development in urban planning [1], intelligent agriculture [2], traffic monitoring [3], disaster management [4], geographical positioning [5], and other aspects. In remote sensing, semantic segmentation is a very important task in image interpretation and processing. The semantic segmentation on high-definition maps usually refers to the recognition of geographical entities (such as woodland, buildings, and water bodies) at the pixel level. Therefore, semantic segmentation is a key procedure to improve understandings of remote sensing images.

The progress of artificial intelligence promotes the application of convolutional neural networks (CNNs) in computer vision [6], [7]. CNN has mighty capacity to voluntarily extract nonlinear and hierarchical features, which significantly affected image processing [8]. Semantic segmentation model is able to perform the task of land cover, largely due to the convolution operation [9], [10]. Traditional models collect local features in a hierarchical manner, enabling the model to have powerful image representations. However, the network is unable to capture global features of the graph [11]. High-resolution remote sensing images produce more complex information, and different sources of image acquisition also cause data interference [12], [13]. Because of the above reasons, the semantic segmentation model cannot pay attention to the key information effectively in the semantic segmentation of land cover. The convolution of the semantic segmentation model cannot enhance nonlocal relationships across the whole image, and it cannot learn complex connections not only between neighbors but also to the neighbors' neighbors (long-range dependence) [14]. Although this problem can be addressed by expanding receptive fields, this may require intensive but disruptive pooling [15]. Therefore, it is difficult to use the existing semantic segmentation models to efficiently and accurately classify land cover from diverse high-resolution remote sensing images.

Transformer originally appeared in natural language processing (NLP). They only employ an attention mechanism to establish connections between tokens in different languages. Transformer quickly dominates the NLP field due to their outstanding performance. Because of the success of NLP, attention mechanisms receive increasing interest in computer vision. At present, the main challenges faced by transformer in semantic segmentation of land cover are as follows: First, visual entities varies greatly, and visual transformer (ViT) does not adapt well to different scenarios. Second, the image has many pixels, and global self-attention makes transformer computationally heavy. The Swin transformer [16] is a general vision architecture proposed to solve these two problems in the semantic

Manuscript received 19 November 2022; revised 28 January 2023, 27 March 2023, and 15 May 2023; accepted 10 July 2023. Date of publication 14 July 2023; date of current version 31 July 2023. This work was supported by the National Natural Science Foundation of China under Grant 42075130. *(Corresponding author: Liguo Weng.)*

Liguo Weng, Kai Pang, and Min Xia are with the Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, B-DAT, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: 002311@nuist.edu.cn; 739389444@qq.com; xiamin@nuist.edu.cn).

Haifeng Lin is with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China (e-mail: haifeng.lin@njfu.edu.cn).

Ming Qian is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: mingqian@whu.edu.cn).

Changjie Zhu is with the Changzhou Campus, Hohai University, Nanjing 210000, China (e-mail: 20030030@hhu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3295729

segmentation of land cover. Unfortunately, although transformer reflects complex spatial transformations and constitutes a global representation, it ignores local feature details, thereby making the partition boundary of the ground entity rough. Therefore, it remains to be a problem that how to accurately blend local and global features.

In this study, a dual-branch parallel network structure, called Sgformer, was designed to solve the problems of semantic segmentation in land cover. We use Transformer blocks in Swin Transformer and CNN as the backbone. On the one hand, transformer has dynamic attention, global context, and better generalization capabilities that are not available in the CNN. On the other hand, CNN has shift, scale, and distortion invariances, which the transformer module lacks. Our proposed method effectively combines the two so that the two branches can take advantage of each other's advantages, which helps the model extract features more efficiently. In terms of feature fusion, the use of feature coupling module (FCM) enables transformer branch and CNN branch to guide each other to carry out feature mining and extract multiscale context information, thus improving the segmentation accuracy of land cover at different scales. In the decoding stage, the features of different levels extracted by the two branches are fully utilized for fusion and decoding, and hence, semantic information and spatial location information are effectively fused. As a result, the location of land cover is more accurate and the partition boundaries are shown in more detail.

Overall, our work had three contributions.

1) A dual-branch parallel network structure called the Sgformer is proposed. It introduces the module design of Swin transformer into the semantic segmentation of land cover and combines it with CNN. Swin transformer extracts the global information and the correlation degree between each pixel and integrates the local features and spatial details extracted by CNN. It effectively complementes the shortcomings of the two in semantic segmentation of land cover and retaines the advantages of each in this task.
2) A semantics guided unit (SGU) is proposed as the basic unit of CNN branch, which can extract deep information of remote sensing image efficiently by integrating spatial details and contextual semantic information. This module greatly reduces the weight of the overall model and greatly improves the efficiency of semantic segmentation.
3) An FCM is proposed, it addresses semantic divergence of features at different levels during fusing. It incorporates not only the local features from the CNN output but also the global representation from the transformer output.
4) A multilevel feature attention upsampling (MFAU) structure is designed, it not only combines high-level features with low-level local features but also uses high-level features to supply guiding information to low-level global representation so as to generate new high-level features during upsampling. This is of great significance for locating and restoring in high-definition remote sensing images.

II. Related Work

A. Traditional Method

Segmentation was of crucial research meaning for vehicle observation [17], land cover mapping [18], change detection [19], building and road extraction [20], and etc. The semantic segmentation technology was gradually used in the field of remote sensing, such as traditional methods (such as logistic regression [21], distance-based measurement [22], and clustering [23]) and machine learning (such as support vector machine [24], random forests [25], artificial neural networks [26], and multilayer perceptrons [27]). The above traditional methods were able to work on small sample images. However, image analysis of large samples had some disadvantages, such as low segmentation accuracy. For example, edge, contour, and texture features were hard to extract, resulting in inadequate reliability [28]. To sum up, the traditional classification on high-definition remote sensing images showed poor feature extraction capability and limited generalization ability. Thus, it was unable to accomplish precise pixel-level classification on high-definition remote sensing images.

B. Convolution-Based Segmentation Models

Early CNN networks were often used for image classification tasks. However, a land cover detection task cannot be regarded as an image classification task, and the image semantic segmentation method was usually used for land cover detection problem [29], [30]. In 2014, a full convolutional network (FCN) [31] was proposed to enable images of any size to be restored to the original size. Deconvolution was used to upsample to the original image. In 2015, SegNet [32] was created to accomplish one-stage pixel-level image classification by building an encoder–decoder symmetrical structure. Cai et al. [33] designed an end-to-end full convolutional network for the characteristics of aerial images and used online samples to process the difference samples during training. However, since high-resolution remote sensing images contain rich ground information, the feature extraction ability of FCN cannot meet more precise semantic segmentation of land cover.

Zhang et al. [34] proposed a new multiscale deep learning model, atrous spatial pyramid pooling (ASPP) Unet, and ResASPP-UNet for urban land cover classification based on very high resolution (VHR) satellite images. ASPP-Unet included a contraction path for extracting high-level features and an extension path for upsampling features to create high-resolution output. An ASPP technique was used in the bottom layer to incorporate multiscale deep features into discriminant features. ResASPP-Unet further improved the architecture by replacing each layer with a Res unit. Nogueira et al. [35] proposed a technology to perform semantic segmentation of remote sensing images. The main idea of this technology is to train atrous convolutional networks with different patch sizes so that they could capture more contextual features from heterogeneous contexts. With the development of semantic segmentation model, encoder–decoder symmetric structure, skip connection structure, ASPP, and so on had greatly improved the accuracy of the model and also promoted the development and application of semantic segmentation in land cover.

C. Transformer

Transformer has recently shown great performance in image processing. Early exploration could be roughly divided into two categories. First, some literature considered attention as a flexible application module that was able to be seamlessly integrated into the existing CNN structures. Representative works included nonlocal networks [36], relation networks [37], and CCNet [38]. Second, some works aimed to replace all convolutional operations with attention mechanisms, such as local relation networks [39] and self-attention. These two directions both showed promising results, but they were still based on CNN architectures.

ViT [40] was a pioneering work using pure transformer architecture in visual recognition. Most of the existing networks in CV were still resnet-like network architectures. Due to the computational efficiency and scalability of the transformer, ViT directly used the standard transformer with minimal changes to cut the map into patches and form a linear embedding sequence to replace the tokens in the original NLP as input for supervised image classification experiments. In the field of semantic segmentation of remote sensing images, Kaselimi et al. [41] proposed ForestViT, a multilabel visual converter model, which takes advantage of the self-attention mechanism, avoiding any convolution operation involved in the common deep learning models used for deforestation detection.

The proposal of pyramid vision transformer (PVT) [42] presented a method to introduce ViT into a CNN-like pyramid structure so that PVT acted as a backbone for dense prediction tasks, such as CNN. Since the VIT itself was a global receptive field, the same Ttransformer encoders were continuously stacked after directly converting the input image into tokens. Therefore, the network computation increased sharply for intensive tasks requiring high-resolution input plus the network directly converted larger patches into tokens, resulting in a larger loss for intensive tasks. Therefore, PVT adopted a similar architecture to CNN and divided the network into different stages. The dimension of each stage was halved compared with the previous stage, which meant that the number of tokens was reduced by four times. Compared with ViT, PVT was able to output feature maps of different scales, which was a huge advantage for semantic segmentation.

The pioneering work of Swin transformer [16] proposed a layered feature representation scheme that demonstrates impressive performance with linear computational complexity. Wang et al. [43] introduced Swin transformer as the backbone of context information extraction and designed a new decoder to restore resolution and generate segmentation map. It was pioneering to inspire researchers in this field to explore the potential and feasibility of Swin transformer more widely in the field of semantic segmentation of land cover. Zhang et al. [44] proposed a deep neural network mixed with Swin transformer and CNN for semantic segmentation of VHR remote sensing images. The model followed the encoder–decoder structure. The encoder module used a new universal backbone Swin transformer to extract features to achieve better long-range spatial dependence modeling. The decoder module drew on some effective blocks and successful strategies of CNN-based models in remote sensing image segmentation.

III. Methodology

With the progress of satellite technology, the increase in resolution of images also causes a dramatic increase in semantic and spatial information. The interference of complex background information makes it difficult to describe land cover, and the wrong extraction of key information meant model failure and decline. The correspondence between local feature and global representation has been studied extensively in the history of semantic segmentation. Local features are compact vector representations of local image neighborhoods. It affects the segmentation of local details in land cover. And as the network deepen, local features will contain rich semantic information, thus improving the network's understanding of different ground entities. Global representation involves complex spatial transformation and long-distance feature dependence. It focuses on key areas in land cover and captured correlations between pixels in remote sensing images.

The extraction of local features and global representation can help the model accurately segment the edge contour of ground entities and resist the interference of surface objects and noises in land cover. In the training process, extracting local features and global representation can make the model learn semantic information from remote sensing images and pay attention to key regions, thus accelerating the convergence of the model. Therefore, how to effectively extract local features and global representations becomes an important issue in improving model segmentation results. Currently, the existing networks are not very skilled in restoring spatial details. As a result, the model for segmentation of land cover still requires improvement in feature extraction and fusion.

This section first introduces the structure of Sgformer, and then elaborated SGU, Swin transformer, FCM, and MFAU modules in detail.

A. Network Architecture

Local features and global representations are extensively studied in computer vision. Local features, which are compact vector representations of local image neighborhoods, constituted an integral part of many computer vision algorithms. Global representation includes contour representation, shape description, object representation over long distance, and so on. In deep learning, CNN collects local features hierarchically through convolution operations and retains local cues as features [45]. Swin transformer is considered to aggregate global representations between compressed patches embedded in a soft manner by cascading MSA modules. Swin transformer uses sliding windows and makes MSA within each window to transfer information between adjacent windows. On the one hand, such

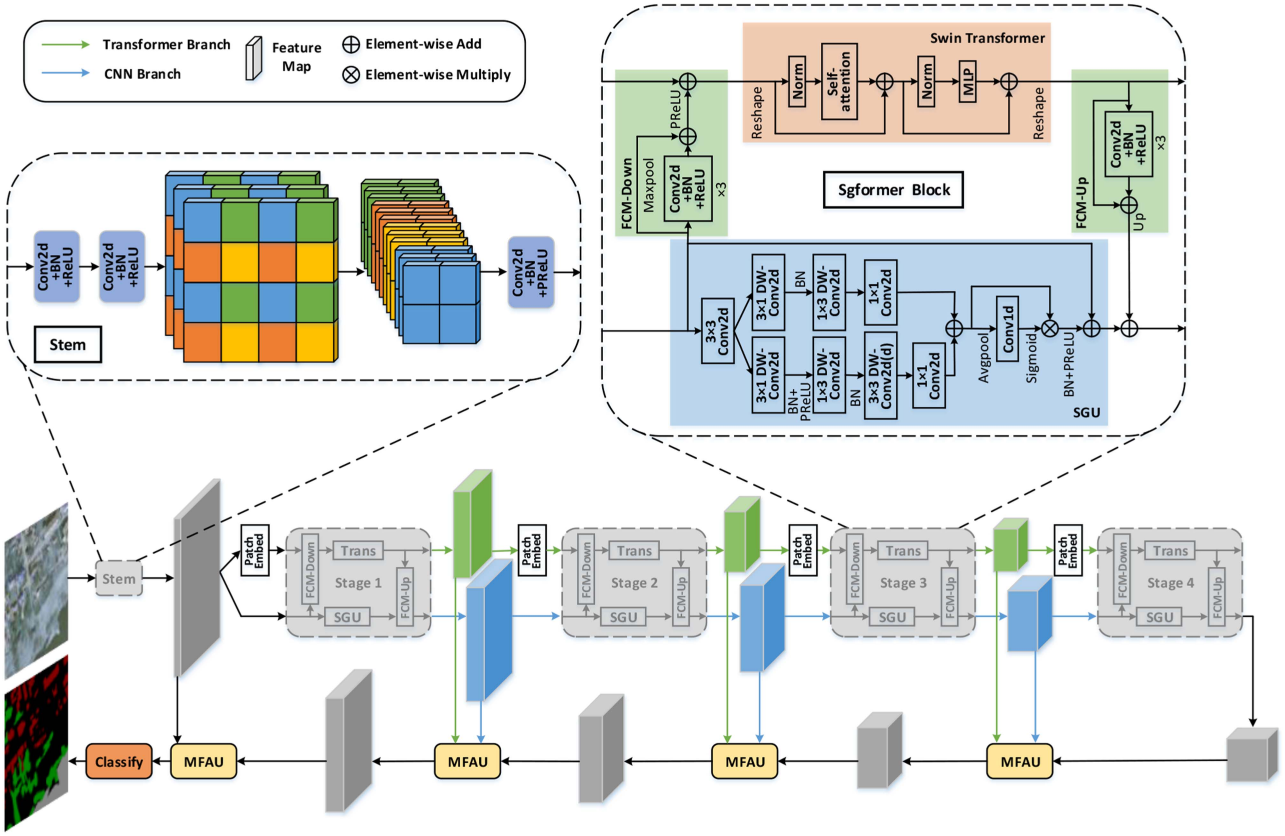

Fig. 1. Structure of Sgformer. Conv1d represents 1-D convolution; Conv2d represents 2-D convolution; DW-Conv2d represents separable 2-D convolution; DW-Conv2d(D) represents separable 2-D dilated convolution.

an operation limited self-attention calculation to each window to reduce computation and, on the other hand, can imitate the localization characteristics of CNN convolution. However, local feature details are still ignored, resulting in a decrease in the background and foreground distinguishability.

Therefore, in order to utilize local and global features, a parallel network architectonics is designed, as shown in Fig. 1, named Sgformer. The network consists of a stem module, a backbone with a dual-branch structure, and multiple MFAU modules. The stem module uses a method similar to adjacent downsampling, which samples every other pixel in a feature image. Consequently, the feature image size is half of the original. At the same time, the channel number is four times of the original. This downsampling mode is similar to the PassThrough layer in YOLO9000 [46]. Traditionally, downsampling is a method of image reduction, which is used to reduce the dimension of features and retained effective information, avoiding overfitting to a certain extent. But that would sacrifice some information. The method of adjacent downsampling can reduce the information loss caused by downsampling and improve the precision of segmentation. This downsampling method ensures that information is not lost. Its structure is shown in Fig. 1.

TABLE I
DETAILED SETTINGS OF SGFOMER BACKBONE

	Layer name	Output size	Settings

	Stem	$\frac{H}{2} \times \frac{W}{2}$	kernel = 3×3; stride = 1;
			padding = 1;

Stage1	CNN anch	$\frac{H}{4} \times \frac{W}{4}$	dim = 64; head = 1;
			depth = 3; win.$sz.$ = 8;

	Transformer branch	$\frac{H}{8} \times \frac{W}{8}$	

Stage2	CNN branch	$\frac{H}{8} \times \frac{W}{8}$	dim = 128; head = 2;
			depth = 4; win.$sz.$ = 8;

	Transformer branch	$\frac{H}{16} \times \frac{W}{16}$	

Stage3	CNN branch	$\frac{H}{16} \times \frac{W}{16}$	dim = 256; head = 4;
			depth = 6; win.$sz.$ = 8;

	Transformer branch	$\frac{H}{32} \times \frac{W}{32}$	

Stage4	CNN branch	$\frac{H}{32} \times \frac{W}{32}$	dim = 512; head = 8;
			depth = 3; win.$sz.$ = 8;

	Transformer branch	$\frac{H}{64} \times \frac{W}{64}$	

Considering the complementation of the two features, we feed global representation from the Swin transformer branch into the CNN to improve the global perception capacity of the CNN branch. Similarly, local features from the CNN branch are fed into the Swin transformer to improve the local perception capacity of the Swin transformer branch. Since Swin transformer has a mass of parameters and calculations, considering the overall learning efficiency and inference speed, we set the structural parameters of Swin transformer in Sgformer, as shown in Table I, which also shows the size of the output feature graphs of each branch and each layer.

B. SGU and Swin Transformer

In semantic segmentation of land cover, the extraction of detailed localization information is of great requirement. Due to the addition of transformer, the calculation amount of the model is very large. Therefore, in the CNN branch, we try to improve the learning efficiency as much as possible and reduce the module weight. We need to maintain a balance between reducing the computational complexity and improving the segmentation accuracy. Using dilated convolutional layers is a very common operation to increase the receptive field, which can make the model more semantically aware. In the conventional lightweight models, the computational complexity of the model is reduced by using depthwise separable convolutions. However, we observe that applying this lightweight operation to the CNN branch will result in a substantial reduction in the overall performance of the model. Therefore, so as to figure out the problem of significant model degradation resulted from lightweight, the unit module SGU of the CNN branch is designed.

SGU adopts ResNet-like residual structure to avoid network degradation caused by the increase of network depth. CNN has the ability to obtain contextual semantic information. And combining different levels of semantic information obtained by CNN can effectively fuse multiscale information. SGU offsets the performance degradation brought by lightweight by simultaneously extracting local spatial detail information and extensive contextual semantic information. A simple 3 × 3 depthwise convolution (DWConv) is used in the embranchment of extracting spatial details to reduce computation. Strip convolution can better obtain detail information, such as contours and boundaries of objects. Therefore, the standard 3 × 3 DWConv is replaced by a 3 × 1 DWConv and a 1 × 3 DWConv. To be able to speed up the learning of the network and prevent gradient disappearing or exploding, a batch normalization (BN) [47] operation is conducted between these two strip convolutions.

After extracting spatial detail information, a 1 × 1 pointwise convolution (PWConv) is used at the bottom of this branch to solve the problem that the DWConv causes the interaction of channel information to be cutoff. To extract contextual semantic information, the method of increasing the receptive field is adopted to effectively perceive the surrounding environment. Therefore, the method of extracting spatial detail information is borrowed, and on this basis, a 3 × 3 dilated DWConv is added between BN and PWConv in the second branch. We adopt a preactivation scheme [48] between strip convolutions and use this scheme after BN. Due to the lightweight design of SGU, using parameterized ReLU (PreLU) as the activation function can achieve better performance than ReLU. Finally, we combine the two pieces of information by addition.

Being able to efficiently transfer information between different channels is crucial for the increase of model performance. Therefore, after extracting spatial detail information and contextual semantic information, an operation that can capture the information correlation between channels is required. Detailed technical details of this operation are in the Appendix.

Because the transformer is a heavy network, its addition with the CNN, which is not designed light weight, will have very low learning efficiency (although good segmentation ability as a heavy network). Therefore, the SGU, actually a lightweight CNN, is designed to achieve better learning efficiency without losing any segmentation ability.

Although it could acquire local features, the model is still incapable in contour representation, shape description, and knowledge retrieval over distant objects. This is because the pieces of information learned by the traditional CNN methods are local features. The global information cannot be truly retrieved by using atrous convolutions to expand receptive field. Therefore, global awareness needs to be addressed by adding a branch to the CNN to extract global representation. See Appendix for detailed technical details.

C. Feature Coupling Module

As the feature of CNN embranchments and transformer embranchments do not match each other's semantic information, an FCM is proposed to couple local and global features by way of interaction.

From Fig. 1, we can see that feature map sizes of the CNN embranchment and the transformer embranchment are inconsistent. The feature image size in CNN is $C \times H \times W$ (C, H, and W are the channel, height, and width, respectively), whereas the feature image in transformer branch is $C \times (H/2) \times (W/2)$. The feature image size of the CNN embranchment is twice that of the transformer embranchment, but their channels are the same. When the feature map is transferred from the CNN branch to the transformer branch, the feature map go through two sub-branches. They are downsampled to the same size using max pooling and cubic convolution operators. The results from these two sub-branches are added to go through a nonlinear activation function (PReLU), and finally are combined with patch embedding.

When the feature graph is transmitted from transformer branch to CNN branch, it is initially output in the form of residual structure by using the triple convolution operator, then it is upsampled to the same size, and finally, it is added to the feature graph of CNN branch. At the same time, we use BatchNorm module and ReLU module to regulate the feature map after each convolution operator.

There are significant semantic differences between the CNN embranchment and the transformer embranchment. This is because the feature maps of the CNN branch are collected through local convolution operators, while the feature maps of the transformer branch are aggregated through a global self-attention mechanism. Therefore, FCM is applied to each block to gradually erase the semantic gap.

D. Multilevel Feature Attention Upsample

The Sgformer designed in this study is a network based on encoder–decoder structure. We gradually complete the fusion and restoration of upsampled features through four MFAU structures. The main methodology is to generate new features after fusing high-level representation with low-level local and global representation. Compared with UNet, this module has less

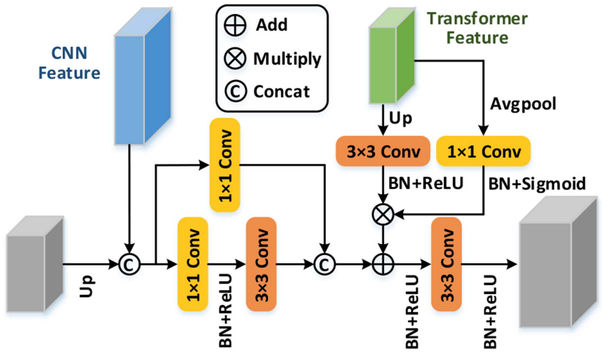

Fig. 2. Structure of MFAU.

parameters and calculation. At the same time, its segmentation is improved, and the high-definition details are better recovered. Different from the two-level feature modules and traditional upsampling modules (such as global attention upsample [49]) and traditional networks (such as Unet), MFAU not only fuses local features from CNN outputs but also fuses global representations from transformer outputs.

Upsampling is applied to recover resolution. This process is very significant. The MFAU structure is used for upsampling on CNN and transformer parallel networks. It effectively utilizes the local information of the CNN branch and the global information of the transformer branch during upsampling so that high-level features can have better perceptual information when restoring resolution. The MFAU module is different from the general upsampling module. It has three inputs (multilevel), two of them are low-level features, which are CNN features and transformer features from the encoding process, and the third one is high-level features. The architecture is revealed in Fig. 2. See Appendix for detailed technical details.

IV. Experiment

A. Datasets

1) Land Cover Dataset: Land cover dataset [50] is applied to examine the performance of Sgformer on land cover segmentation. This dataset consisted of remote sensing images selected from aerial photographs. It is manually labeled into four categories: building, woodland, water, and background. The images in the dataset are cropped into 512 × 512 images. Then, these images are processed for data enhancement with three strategies: horizontal flip (50%), vertical flip (50%), and random rotation (-10° to 10°). Data augmentation not only expands the dataset but also increases the disturbance during model training. At the same time, the generalization ability of the model is enhanced. The cropped image and its labels were shown in Fig. 3. This article uses holdout verification method and divides the dataset according to the current mainstream training/validation set separation ratio of 80%: 20%. Take the average of three experiments as the final result.

The dataset has the following characteristics.

1) It contains objects similar to buildings, such as containers and large trucks, plus the tops of some houses in the dataset are similar to water bodies and backgrounds, and the detection ability of the model can be well examined.
2) It includes many scenes covered, including farmland, roads, factories, houses, and etc. Therefore, segmentation ability of the model can be truly evaluated.
3) Water category includes running water and standing water but excluded ditches and dry riverbeds. These maps may result in apparent indistinguishability. The object surrounded by the yellow rectangle in Fig. 3(a) looks like the background, but it is actually water.
4) Woodland consists of a group of trees, not individual trees and orchards. The object enclosed by the yellow rectangle in Fig. 3(b) looks like woodland, but it is actually a shrub and should be identified as the background. Overall, the accurate land cover classification is hard on this dataset.

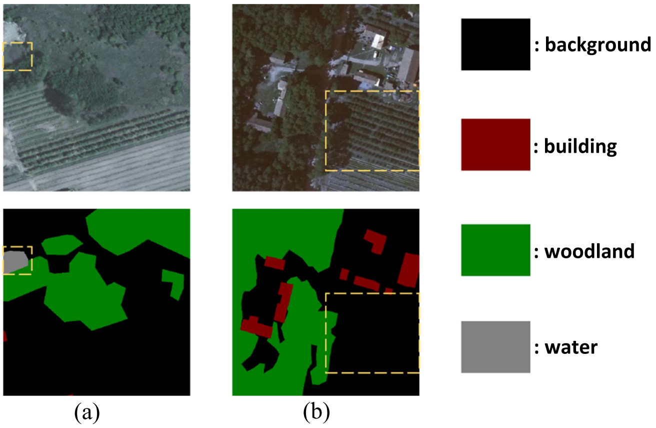

Fig. 3. Images showing a part of the land cover dataset. (a) First sample. (b) Second sample.

2) Inria Aerial Image Labeling Dataset (IAILD): The dataset is an IAILD [51]. IAILD addresses a central subject in remote sensing: automatic pixel labeling of aerial images. It has two semantic classes of ground truth data: build and nonbuild. The images cover different urban areas. The dataset is manually labeled into two categories: building and background. We crop the image to 256 × 256 to examine the performance of the model at different scales. The holdout verification method is used, and the training/validation set separation ratio is divided 80%:20%. Take the average of three experiments as the final result. The cropped image and its labels are shown in Fig. 4. As shown in Fig. 4(a), objects, such as containers, are very similar to buildings which we want to identify; therefore, false detections can happen easily. From Fig. 4(b), it can be found that trees occlude buildings, which make it difficult to segment the outlines of buildings. Therefore, this dataset can examine the detection capacity of the network while testing its generalization ability.

B. Implementation Details

All experiments in this study were run on the Ubuntu (Version: 5.4.0-6ubuntu1) system and done on the Pytorch [52] framework. The hardware platform used in this article is composed of Intel Xeon E5-2678 v3 @ 2.50 GHz CPU and NVIDIA RTX2080Ti GPU. We choose the adaptive moment (Adam) [53]

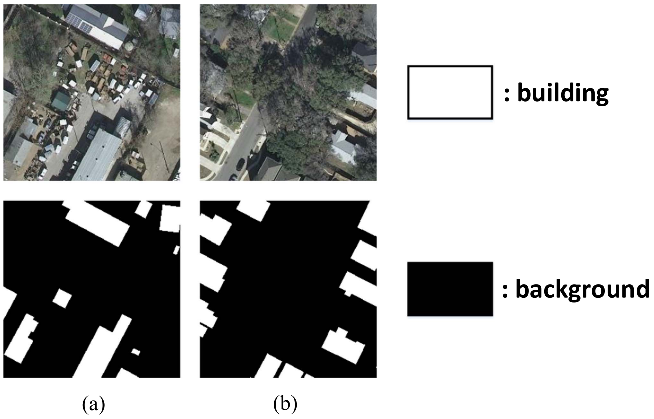

Fig. 4. Images showing a part of IAILD. (a) First sample. (b) Second sample.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODULES IN THE MODEL

Method	PA(%)	MPA(%)	MIOU(%)	FPS	Params(M)
Swin	86.07	87.57	65.32	56.06	20.75
SGU	94.17	91.23	84.27	79.58	12.65
Swin+SGU	94.32	91.73	85.01	50.72	28.58
Swin+SGU+FCM	94.51	93.26	85.48	49.34	30.45
Swin+SGU+FCM+MFAU	94.32	92.86	85.82	48.30	32.22

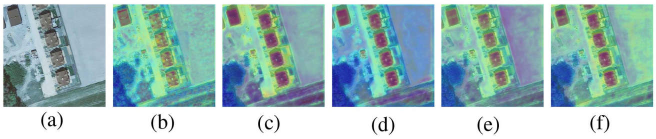

Fig. 5. Visual comparison of different modules. (a) Real image. (b) Swin. (c) SGU. (d) Swin+SGU. (e) Swin+SGU+FCM. (f) Swin+SGU+FCM+MFAU.

optimizer to optimize the model when training the model. In the Adam, the rate β_1 is set to a factor of 0.9, and the rate β_2 is set to a factor of 0.999. When training the data, the initial learning rate of the model was 0.001. In order to avoid overfitting, L_2 regularization with a value of 0.0001 in the Adam optimizer is introduced. When setting the learning rate, this article adopts the "poly" learning rate strategy, that is, the current learning rate is equal to the initial learning rate multiplied by $1-(\frac{\text{iter}}{\text{max iter}})^{\text{power}}$, where iter represents the number of rounds of the current iteration, and max iter represents the total number of iterative rounds during training. The exponent power is 0.9 in this article, and the max iter is 600. The total number of epochs is 300. Since we only use a single Nvidia RTX2080Ti with only 11 Gb of memory, the training batch size is 4. The cross entropy is used as the loss function in this article, and its expression is given as follows:

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \quad (1)$$

where N represents the number of samples, M represents the number of categories, y_{ic} represents the sign function (0 or 1), and p_{ic} represents the predicted probability that the observed sample i belongs to category c.

C. Ablation Study on Land Cover Dataset

In this section, ablation experiments are done stepwise to prove the impact of each module on the land cover dataset. For evaluation, we mainly use pixel accuracy (PA), mean pixel accuracy (MPA), and mean intersection over union (MIOU), as well as frames per second (FPS) and params to observe the reasoning speed and computation of the model, and we visualize the network after each addition of modules. Therefore, this article defines a new score map, as shown in the following equation:

$$P_2 = P_1 + M(P_1) \quad (2)$$

where P_1 is a coarse score map, P_2 is a rerepresented finer score map, and M represents our newly added module. The visual operation is inspired by the fact that humans focus on different positions when looking at objects. Through the visualization operation, we can intuitively see that the network paid more attention to region of the desired category. Therefore, areas with a high degree of attention have higher scores in the score map, and the color of the corresponding area will be darkened.

In the following experiments, we first use the CNN branch and the transformer branch as the basic feature extraction network, respectively. And in the decoding part, the same decoding method as UNet is used to splice the previous feature image and the upsampled feature image together by a skip connection. The parameter settings of Swin transformer remain consistent with the Sgformer proposed in this article, as shown in Table I for details. Through experiments on the land cover dataset, it can be found from the MIOU in Table II that the results of Swin using only transformer and of SGU using only CNN are not satisfactory. Especially, for Swin transformer, when the network is designed lightweight, which has superiority in parameter number and computation cost, its ability to extract useful features drops significantly.

From the visual diagrams in Fig. 5(b) and (c), it can be intuitively seen that Swin transformer can better retrieve global information through its advantages in self-attention so as to accurately focus on key areas. But it has serious defects in the details. As reflected in Fig. 5(b), both buildings and woodlands are accurately focused, but the outlines of buildings and woodlands are not clear. On the contrary, SGU is the opposite. Although local details of building and woodland are accurately extracted, SGU cannot effectively distinguish woodland and grassland due to the lack of global representation. Therefore, we combine Swin transformer with SGU, but we used direct connections for the two branches.

Through the MIOU in Table II and the visual map in Fig. 5(d), it is not difficult to see that the result of the network is built up. As reflected in Fig. 5(d), the contour of buildings and woodland is clear, while the attention on grassland area is reduced. However, since the feature map in SGU and the patch embedding in Swin transformer do not match each other, a direct connection will lead to a repulsion phenomenon in the process of fusing features

TABLE III
COMPARISON OF DIFFERENT MODEL EVALUATION METRICS ON LAND COVER DATASET

Method	Building			Woodland			Water			PA(%)	MPA(%)	MIOU(%)	FPS	Params(M)
	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$(%)					
BiSeNetV2	84.65(4)	75.65(10)	79.90(7)	90.58(14)	94.07(6)	92.30(10)	96.49(5)	93.30(11)	94.87(9)	93.36(9)	91.68(8)	82.84(10)	**125.21(1)**	3.62(4)
DeepLabV3+	83.15(10)	77.72(7)	80.34(6)	91.46(10)	94.25(4)	92.83(7)	96.80(4)	94.09(6)	95.43(6)	93.84(6)	91.66(9)	83.67(8)	42.60(8)	59.34(13)
DenseASPP	79.11(14)	80.00(2)	79.55(8)	**93.73(1)**	92.53(11)	93.13(4)	96.28(8)	95.33(2)	95.80(2)	94.15(2)	90.88(13)	83.84(5)	90.66(4)	2.48(3)
DFANet	84.20(6)	71.62(12)	77.40(11)	92.23(4)	92.25(12)	92.24(11)	95.52(13)	93.54(10)	94.52(11)	93.30(10)	91.44(10)	81.80(11)	40.43(10)	2.18(2)
ExtremeC3	81.89(12)	76.81(8)	79.27(9)	92.22(5)	93.34(7)	92.78(8)	95.98(10)	94.08(7)	95.02(8)	93.76(7)	91.20(12)	83.07(9)	105.32(2)	**0.04(1)**
GhostNet	72.52(16)	63.53(15)	67.73(16)	91.08(12)	91.63(14)	91.35(14)	95.46(14)	91.28(13)	93.32(13)	92.30(14)	88.02(16)	77.52(14)	70.36(5)	8.56(5)
HRNet	86.27(3)	79.11(4)	82.54(2)	91.90(9)	92.99(9)	92.44(9)	95.81(12)	93.91(9)	94.85(10)	93.57(8)	92.12(3)	83.93(4)	21.35(15)	65.85(14)
OCRNet	83.26(9)	78.88(5)	81.01(4)	90.95(13)	**95.09(1)**	92.98(5)	97.02(3)	94.14(5)	95.56(4)	93.94(4)	91.76(6)	84.06(3)	19.64(16)	70.35(15)
PAN	83.11(11)	77.84(6)	80.39(5)	91.35(11)	94.47(3)	92.89(6)	97.12(2)	93.95(8)	95.51(5)	93.87(5)	91.73(7)	83.75(7)	92.48(3)	23.69(7)
PSPNet	83.87(8)	76.05(9)	76.77(12)	92.17(6)	94.18(5)	93.16(3)	96.12(9)	95.26(3)	95.69(3)	94.10(3)	91.86(5)	83.83(6)	28.88(13)	49.07(12)
UNet	**87.33(1)**	72.42(11)	79.18(10)	93.19(2)	90.35(15)	91.74(13)	96.29(7)	88.26(15)	92.10(14)	92.78(11)	92.25(2)	80.96(12)	33.67(12)	17.27(6)
SegNet	83.92(7)	58.94(16)	69.24(15)	92.14(7)	91.88(13)	92.01(12)	**97.22(1)**	83.04(16)	89.57(16)	92.40(12)	91.35(11)	76.71(16)	36.38(11)	29.45(9)
PVT	80.52(13)	66.69(13)	72.95(13)	88.54(16)	89.66(16)	89.10(16)	94.21(16)	89.11(14)	91.59(15)	90.60(15)	88.73(15)	76.76(15)	51.67(6)	25.96(8)
Swin-UNet	78.89(15)	65.76(14)	71.73(14)	89.89(15)	92.65(10)	91.25(15)	95.18(15)	91.80(12)	93.46(12)	92.31(13)	89.43(14)	78.71(13)	42.31(9)	41.39(11)
TransUNet	84.33(5)	79.12(3)	81.64(3)	93.15(3)	93.24(8)	93.20(2)	95.85(11)	94.52(4)	95.18(7)	94.15(2)	92.02(4)	84.31(2)	22.13(14)	105.91(16)
Sgformer(our)	87.28(2)	**82.04(1)**	**84.58(1)**	92.05(8)	94.55(2)	**93.28(1)**	96.39(6)	**95.48(1)**	**95.94(1)**	**94.32(1)**	**92.86(1)**	**85.82(1)**	48.30(7)	32.22(10)

The bold entities indicate the best results.

so that the final result did not meet our expectations. To solve this problem, we used FCM at the interface of Swin transformer and SGU to replace the direct connection to eliminate the semantic divergence between them. The MIOU in Table II and the visual map in Fig. 5(e) show that FCM can further improve the MIOU of the model, and FCM improves the performance by nearly 0.47%.

In order to recover high-definition details better, we replace the original encoder with MFAU. In the upsampling process, MFAU not only fuses local features of the CNN branch but also adds the representation of the transformer branch. MFAU enables the model to better locate the spatial position of each category of objects when restoring high-definition maps, and it further improves the model accuracy. Its performance is shown in Table II and Fig. 5(f). While accurately identifying building and woodland, it can also capture their detailed information.

D. Comparative Experiment With Other Networks

1) Comparative Experiment on Land Cover Dataset: So as to efficiently prove that the methodology designed in this study is able to well and truly segment buildings, woodlands, and water bodies in different scenarios, the method is contrasted with other models on land cover dataset. The other semantic segmentation networks are the traditional CNN models, such as BiSeNetV2 [54], DeepLabV3+ [55], DenseASPP [56], DFANet [57], ExtremeC3 [58], GhostNet [59], HRNet [60], OCRNet [61], PAN, PSPNet [62], UNet, and SegNet. We also compare the performance with models using transformer, such as PVT [42], Swin-UNet [63], and TransUNet [64]. These methods are currently relatively excellent deep learning methods.

Table III presents the comparison outcomes of the networks in the same experimental environment. It can be seen that our designed methodology is meaningfully higher than other networks on MIOU, with at least 1.51% improvement. Our method also achieves most of the lead on different classes of precision (P), recall (R), and $F1$. Wilcoxon rank-sum test shows that Sgformer's performance is better than the other models. So as to more intuitively show the accuracy of our method on land cover, semantic segmentation ability of the network is evaluated from two aspects: category detection and contour detection.

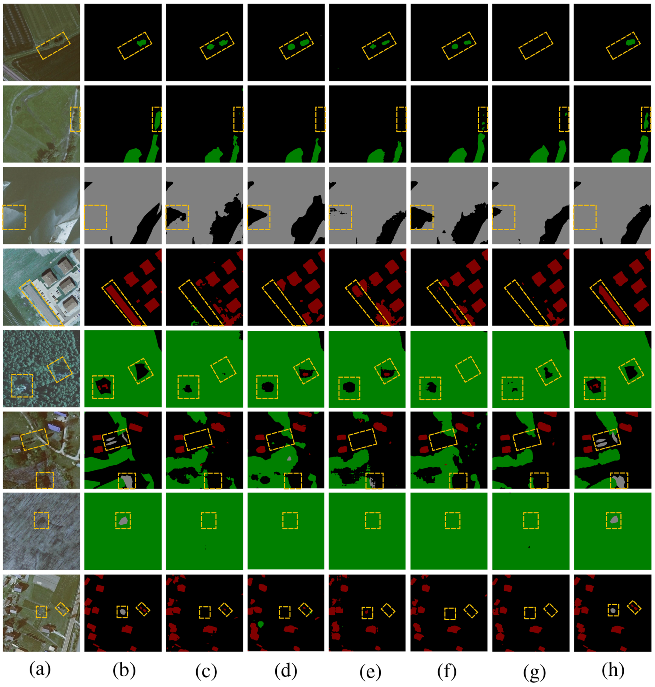

Fig. 6. Comparative experiments on category detection. (a) Real image. (b) Label. (c) DeepLabV3+. (d) BiSeNetV2. (e) DenseASPP. (f) PAN. (g) TransUNet. (h) Our segmentation results.

Fig. 6 shows the experimental results of category detection. A single tree can not constitute woodland, and two or more trees

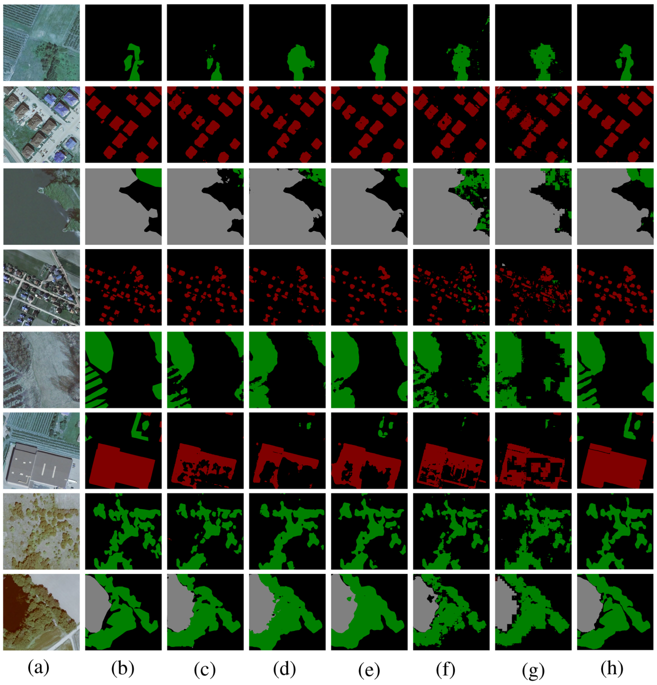

Fig. 7. Comparative experiments in contour detection. (a) Real image. (b) Label. (c) DFANet. (d) UNet. (e) PSPNet. (f) PVT. (g) Swin-UNet. (h) Our segmentation results.

TABLE IV
COMPARISON OF EVALUATION METRICS OF DIFFERENT MODELS ON IAILD

Method	Building			PA(%)	MPA(%)	MIOU(%)
	P(%)	R(%)	$F1$(%)			
BiSeNetV2	87.63(4)	83.14(11)	85.33(9)	94.27(9)	91.73(5)	83.77(9)
DeepLabV3+	**88.45(1)**	83.04(12)	85.66(7)	94.43(3)	92.13(2)	84.12(5)
DenseASPP	87.59(5)	83.82(8)	85.66(7)	94.38(6)	91.79(4)	84.08(7)
DFANet	84.91(14)	80.75(13)	82.78(12)	93.27(13)	90.07(13)	81.29(12)
ExtremeC3	86.13(10)	83.70(9)	84.90(11)	94.04(12)	91.04(11)	83.30(11)
GhostNet	84.93(13)	78.96(14)	81.84(13)	92.98(14)	89.88(14)	80.46(13)
HRNet	86.18(9)	85.35(4)	85.76(6)	94.33(7)	91.26(9)	84.11(6)
OCRNet	87.67(3)	84.04(7)	85.82(2)	94.44(2)	91.86(3)	84.24(2)
PAN	87.23(7)	84.39(5)	85.78(4)	94.40(5)	91.68(7)	84.18(4)
PSPNet	87.33(6)	84.30(6)	85.79(3)	94.41(4)	91.72(6)	84.19(3)
UNet	85.62(12)	85.44(3)	85.53(8)	94.21(10)	90.99(12)	83.87(8)
SegNet	87.09(8)	83.34(10)	85.18(10)	94.19(11)	91.48(8)	83.60(10)
PVT	84.67(15)	74.63(16)	79.33(15)	92.21(15)	89.25(15)	78.30(15)
Swin-UNet	82.43(16)	76.96(15)	79.60(14)	92.10(16)	88.38(16)	78.38(14)
TransUNet	85.99(11)	85.54(2)	85.77(5)	94.31(10)	91.19(10)	84.11(6)
Sgformer(our)	88.17(2)	**85.86(1)**	**87.00(1)**	**94.86(1)**	**92.33(1)**	**85.39(1)**

The bold entities indicate the best results.

are required to be considered woodland. Therefore, most of the networks cannot distinguish a single tree from woodland, as shown in the first and second rows of Fig. 6. The third and fourth rows of Fig. 6 show that other models cannot accurately detect water bodies and buildings, while our model can accurately identify both categories. This is due to the addition of SGU's semantic perception, which enables the model to accurately classify pixels. As shown in the fifth line of Fig. 6, tree shade interfered with house detection, which tests the extraction ability of semantic information and spatial location information of the model. Similarly, the sixth and seventh lines of Fig. 6 show that the shade of trees is similar to the color of water, which make it impossible for other models to effectively identify, resulting in false detection and missing detection.

The proposed method uses CNN branch and transformer branch to extract semantic information and spatial location information, respectively, which reduce the problem of insufficient extraction of semantic information and spatial location information in the existing models. The eighth line of Fig. 6 shows the interference of noise to the recognition of small objects and objects similar to the background. We can see that other methods for noise interference suppression are insufficient. The method in this article benefits from the effective fusion of local feature and global representation so that the object can still be detected even with noise interference.

Fig. 7 shows the experimental results of contour detection. As buildings and trees are shaded by light, this creates a challenge to separate buildings from woodlands. The model needs to distinguish itself from the shadow effectively to make the segmentation boundary fine. Small objects around buildings, woodland, and water will also affect the recognition of object contour features. As shown in the first, second, fifth, seventh, and eighth lines of Fig. 7, Sgformer can efficiently avert the disturbance of other entities to segment buildings, woodlands, and water bodies. This benefits from the advantages of transformer in contour representation, shape description, and long-distance feature dependence.

Sgformer has both global information extracted by transformer and local information extracted by CNN, and the two can complement and guide each other. The semantic divergence between them is eliminated by FCM. High-level local features and global representations can also be effectively utilized through MFAU when recovering the position of each pixel. Therefore, the shape recognition of small objects and contour detection of large objects put forward higher requirements for model detection ability. As shown in the third, fourth, and sixth lines of Fig. 7, large areas of water and buildings and small areas of woodland and houses all expose the problem of severe loss of boundary details in other models. The model in this article fully extracts multiscale context information through CNN branch and transformer branch, and integrates the features of the two branches so that the model can accurately locate buildings, woodland, and water.

2) Comparative Experiment on IAILD: So as to examine the generalization performance of our network, the experiment is done on IAILD. The dataset also evaluates how well the model detected complex backgrounds and details. We select the models tested on the land cover dataset for comparison. The performance is seen in Table IV. Table IV presents the comparison outcomes of diverse models in the same experimental environment. It can be seen that our designed methodology is obviously higher than other deep learning models on MIOU, with at least 1.15% improvement. Our method also achieves most of the lead on different classes of P, R, and $F1$. Wilcoxon rank-sum test shows that Sgformer's performance is better than other models.

The visualized results on the test set are shown in Fig. 8. The first and second rows of the consequences show the great

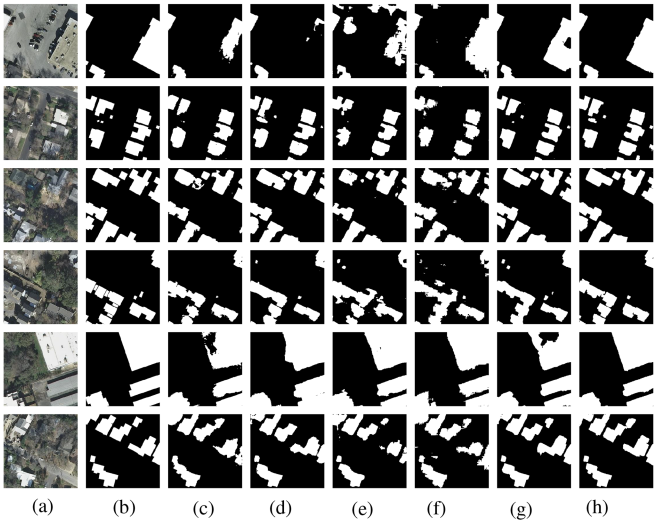

Fig. 8. Comparative experiments of different models on IAILD. (a) Real image. (b) Label. (c) GhostNet. (d) ExtremeC3. (e) PVT. (f) Swin-UNet. (g) TransUNet. (h) Our segmentation results.

segmentation of buildings by Sgformer. At the same time, the performance in the third and fourth rows shows that Sgformer removes the background interference. The outcomes in the fifth and sixth rows show relatively complete predictions for buildings compared with baseline networks (GhostNet, ExtremeC3, PVT, Swin-UNet, and TransUNet). These results are consistent with the predictions from the land cover dataset, where Sgformer performs the best. These results once again prove that the network has a powerful capability of feature capture and a strong capacity of high-definition detail recovery.

V. Discussion

The progress of CNN has promoted the development of semantic segmentation in land cover, but it is largely attributed to the convolutional operation. Therefore, the traditional semantic segmentation model has strong local feature extraction ability in land cover task. And with the increase of network depth, multilevel convolution operation enables the model to learn semantic information in land cover. The advantages of CNN mentioned above are reflected in fine boundary segmentation and semantic understanding of ground entities. However, because high-resolution remote sensing images generate more complex information and different image acquisition sources also cause data interference, the traditional semantic segmentation models cannot effectively focus on key information, making it difficult to greatly improve segmentation accuracy. The application of transformer in vision provides us with new ideas to solve the above problems. Transformer can effectively make up for the defect of CNN in semantic segmentation of land cover due to its ability to extract global representation and master the correlation between long-distance visual elements. Therefore, we want to solve the problem of semantic segmentation on land cover by designing a network that can extract local features and global representation.

In this article, the network adopts double branch design, extracting local feature and global representation, respectively. The CNN branch for extracting local features is composed of SGU. SGU is lightweight in design and can extract local spatial details and extensive contextual semantic information. The transformer branch that extracts the global representation uses the Swin transformer design. This branch improves the network's ability in global perception to focus on key areas in land cover and to capture the correlation between pixels in remote sensing images. Due to the feature mismatch between CNN and transformer, we design FCM as a bridge to eliminate semantic divergence between features at different levels. This article proposes a network based on encoder–decoder structure. Therefore, we need multiple MFAU modules to gradually complete the upsampling feature fusion and recovery. In addition, MFAU module effectively makes use of the local features of the CNN branch and the global representation of the transformer branch in the upsampling process, which make the high-level features have better perception information when restoring the resolution. The combination of SGU, Swin transformer, FCM, and MFAU constitutes the network in this article, which is used to reduce the probability of false detection and missing detection in semantic segmentation of land cover, and at the same time produced clear segmentation boundaries.

Although Sgformer proposed in this article has outstanding performance in semantic segmentation of land cover, compared with other models, Sgformer is deficient in the number of parameters and inference speed due to the transformer structure. Therefore, it takes a lot of time in the training process. In the next improvement, we will make lightweight design for transformer, which cannot only improve the network performance but also reduce the network computing complexity.

VI. Conclusion

Semantic segmentation of land cover is of great application value in high-definition remote sensing images. In this study, a dual-branch parallel network structure is designed, called Sgformer. Sgformer has improvement in two areas: deep feature capture and upsampling feature integration. CNN collects local features in a hierarchical manner, enabling the network to have powerful image representations. But CNN cannot extract global features well. Transformer obtains long-range feature dependence through global self-attention. However, transformer has good and bad performance in different scenarios, and inference speed is also limited. Therefore, we use SGU to extract local features, Swin transformer to extract global representation, and FCM to couple the two semantically divergent features. The fusion of local and global features greatly enhances representation learning of the network and improves the depth feature capture of the network. MFAU is applied to progressively improve the recovered high-definition feature images. After upsampling high-level features, we fuse low-level local features and global representations so that the decoder has better fixed pixel localization ability. On land cover dataset, experiments show that the methodology achieves 85.82% MIOU. The generalization ability of the model is great, obtaining an MIOU of 85.39% on IAILD.

Appendix

A. Semantics Guided Unit

First, we aggregate the information between each channel into a weight that can measure the importance of different channels to accurately map features. This process can be represented as follows:

$$\mathrm{Avg}(c) = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j)}{H \times W}. \tag{3}$$

This equation can be understood as the global average information for the *c*th channel. When the weights between each channel are obtained, the spatial correlation is decoupled. And next, the correlation between channels is captured by one-dimensional (1-D) convolution. This process can be represented by

$$w_c = \mathrm{C1D}(\mathrm{Avg}) \tag{4}$$

where C1D represents the 1-D convolution. We set the convolution kernel of 1-D convolution to 3. w_c is the *c*th component of the weight vector after 1-D convolution. Each component represents the feature map in each channel to capture the interaction information between surrounding channels. The weight vector with the interaction information between the surrounding channels is passed through the Sigmoid function to acquire the weight of each channel. At last, these joint features are reweighted to focus on the importance between different channels. This process can be represented by

$$z_c = \mathrm{Sigmoid}(w_c) \cdot x_c(i,j) \tag{5}$$

$$\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{6}$$

the Sigmoid function is represented by (6).

B. Swin Transformer

Compared with the traditional MSA module, Swin transformer is built around shifted windows. Each Swin transformer block is made up of layer norm (LN), MSA module, residual connection structure, and multilayer perceptron. A windows MSA module and a shifted windows MSA module are applied on coherent Sgformer blocks, respectively. For tokenization, the feature map from the previous layer is compressed into patch embedding through a linear projection layer without overlap. This projection uses a 2 × 2 convolution with a stride of 2. Based on this windowing mechanism, the Swin transformer block can be represented by the following equations:

$$z_i' = \mathrm{MSA}(\mathrm{LN}(z_{i-1})) + z_{i-1} \tag{7}$$

$$z_i = \mathrm{MLP}(\mathrm{LN}(z_i')) + z_i' \tag{8}$$

where LN stands for the layer normalization operator; and z_i is the encoded image representation. Relative displacement bias $B \in R^{M^2 \times M^2}$ is introduced for each head in self-attention

$$\mathrm{Attention}(Q,K,V) = \mathrm{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{9}$$

where $Q, K,$ and V are the query, key, and value matrices; d is the dimension of Q/K; and M^2 is the number of patches in a window.

C. Multilevel Feature Attention Upsample

High-level features are upsampled and concatenated with CNN features. The concatenated feature maps pass through a dual-branch structure. Both of these two branches use a 1 × 1 Conv to compress the channel into half of the original one, which can pass information between channels and reduce the amount of calculation. A 3 × 3 Conv is used in one of the branches as a spatial detail extractor, and the other branch retained the original information. Next, the output outcomes of the branches are spliced. At this point, the amalgamation of high-level information to low-level local information is completed. Since the backbone network also contained low-level global features, a structure that effectively fused global information is designed. Since the size of the transformer feature image is half of the CNN one, we upsample the transformer features and use a 3 × 3 Conv as feature coupler. In order to focus on the significance of channels after the transformer, global context information is focused on by passing the transformer feature through the global average pooling layer. Next, the weight vectors of the channel number with the transformer feature length are obtained.

To extract the correlation between channels, the 1 × 1 Conv bottleneck is applied to deal with the features of the whole channel. Then, the weight vectors are weighted to the transformer features to increase the focus on effective feature channels. Because the feature size of this weight vector is $C \times 1 \times 1$, which is different from the sample size of the transformer feature, the weight vector cannot be directly matrix dot product with the sampled feature of transformer. Therefore, the weight vector of size $C \times 1 \times 1$ is manipulated by the broadcast mechanism to be the same size as the sample size on the transfomer feature. After feature mapping, the transformer features are added with high-level features that fuse local information, and the added result is passed through a 3 × 3 Conv to blend the transformer features.

References

[1] B. Chen, M. Xia, M. Qian, and J. Huang, “MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images,” *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 5874–5894, 2022.

[2] M. T. Chiu et al., “Agriculture-vision: A large aerial image database for agricultural pattern analysis,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2825–2835.

[3] S. Ø. Larsen, A.-B. Salberg, and L. Eikvil, “Automatic system for operational traffic monitoring using very-high-resolution satellite imagery,” *Int. J. Remote Sens.*, vol. 34, no. 13, pp. 4850–4870, 2013.

[4] B. J. Wheeler and H. A. Karimi, “Deep learning-enabled semantic inference of individual building damage magnitude from satellite images,” *Algorithms*, vol. 13, no. 8, 2020, Art. no. 195.

[5] S. Hu and G. H. Lee, “Image-based geo-localization using satellite imagery,” *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1205–1219, 2020.

[6] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, and B. Chen, “Axial cross attention meets CNN: Bibranch fusion network for change detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 32–43, Nov. 2023.

[7] C. Lu, M. Xia, and H. Lin, “Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation,” *Neural Comput. Appl.*, vol. 34, pp. 6149–6162, 2022.

[8] Y. Qu, M. Xia, and Y. Zhang, “Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow,” *Comput. Geosci.*, vol. 157, 2021, Art. no. 104940.

[9] J. Gao, L. Weng, M. Xia, and H. Lin, "MLNet: Multichannel feature fusion Lozenge network for land segmentation," *J. Appl. Remote Sens.*, vol. 16, no. 1, 2022, Art. no. 016513.

[10] S. Miao, M. Xia, M. Qian, Y. Zhang, J. Liu, and H. Lin, "Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery," *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 5940–5960, 2022.

[11] L. Weng and S. Zhang, "STPGTN–A multi-branch parameters identification method considering spatial constraints and transient measurement data," *Comput. Model. Eng. Sci.*, vol. 136, no. 3, pp. 2635–2654, 2023.

[12] Z. Ma, M. Xia, L. Weng, and H. Lin, "Local feature search network for building and water segmentation of remote sensing image," *Sustainability*, vol. 15, no. 4, 2023, Art. no. 3034.

[13] K. Hu, E. Zhang, M. Xia, L. Weng, and H. Lin, "MCANet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images," *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 1055.

[14] K. Hu, M. Li, M. Xia, and H. Lin, "Multi-scale feature aggregation network for water area segmentation," *Remote Sens.*, vol. 14, no. 1, 2022, Art. no. 206.

[15] K. Hu, Y. Ding, J. Jin, L. Weng, and M. Xia, "Skeleton motion recognition based on multi-scale deep spatio-temporal features," *Appl. Sci.*, vol. 12, no. 3, 2022, Art. no. 1028.

[16] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[17] N. Audebert, B. Le Saux, and S. Lefèvre, "Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images," *Remote Sens.*, vol. 9, no. 4, 2017, Art. no. 368.

[18] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multiscale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5609519.

[19] D. Wang, L. Weng, M. Xia, and H. Lin, "MBCNet: Multi-branch collaborative change-detection network based on Siamese structure," *Remote Sens.*, vol. 15, no. 9, 2023, Art. no. 2237.

[20] J. Chen, M. Xia, D. Wang, and H. Lin, "Double branch parallel network for segmentation of buildings and waters in remote sensing images," *Remote Sens.*, vol. 15, no. 6, 2023, Art. no. 1536.

[21] G. N. Rutherford, A. Guisan, and N. E. Zimmermann, "Evaluating sampling strategies and logistic regression methods for modelling complex land cover changes," *J. Appl. Ecol.*, vol. 44, no. 2, pp. 414–424, 2007.

[22] Q. Du and C.-I. Chang, "A linear constrained distance-based discriminant analysis for hyperspectral image classification," *Pattern Recognit.*, vol. 34, no. 2, pp. 361–373, 2001.

[23] U. Maulik and I. Saha, "Automatic fuzzy clustering using modified differential evolution for image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3503–3510, Sep. 2010.

[24] Y. Guo, X. Jia, and D. Paull, "Effective sequential classifier training for SVM-based multitemporal remote sensing image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3036–3048, Jun. 2018.

[25] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[26] C. Adede, R. Oboko, P. W. Wagacha, and C. Atzberger, "A mixed model approach to vegetation condition prediction using artificial neural networks (ANN): Case of Kenya's operational drought monitoring," *Remote Sens.*, vol. 11, no. 9, 2019, Art. no. 1099.

[27] H. Li, C. Zhang, S. Zhang, and P. M. Atkinson, "A hybrid OSVM-OCNN method for crop classification from fine spatial resolution remotely sensed imagery," *Remote Sens.*, vol. 11, no. 20, 2019, Art. no. 2370.

[28] C. Lu, M. Xia, M. Qian, and B. Chen, "Dual-branch network for cloud and cloud shadow segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5410012.

[29] C. Zhang, L. Weng, L. Ding, M. Xia, and H. Lin, "CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery," *Remote Sens.*, vol. 15, no. 6, 2023, Art. no. 1664.

[30] Z. Ma, M. Xia, H. Lin, M. Qian, and Y. Zhang, "FENet: Feature enhancement network for land cover classification," *Int. J. Remote Sens.*, vol. 44, no. 5, pp. 1702–1725, 2023.

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

[32] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[33] B. Cai, Z. Jiang, H. Zhang, Y. Yao, and S. Nie, "Online exemplar-based fully convolutional network for aircraft detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 7, pp. 1095–1099, Jul. 2018.

[34] P. Zhang, Y. Ke, Z. Zhang, M. Wang, P. Li, and S. Zhang, "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery," *Sensors*, vol. 18, no. 11, 2018, Art. no. 3717.

[35] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.

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

[37] A. Santoro et al., "A simple neural network module for relational reasoning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 4974–4983.

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

[39] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3463–3472.

[40] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[41] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023.

[42] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 548–558. .

[43] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 6506105.

[44] 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, Jan. 2022, Art. no. 4408820.

[45] Z. Wang, M. Xia, M. Lu, L. Pan, and J. Liu, "Parameter identification in power transmission systems based on graph convolution network," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 3155–3163, Aug. 2022.

[46] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[49] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*.

[50] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1102–1110.

[51] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[52] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–4.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

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

[55] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Semantic scene segmentation in unstructured environment with modified DeepLabV3+," *Pattern Recognit. Lett.*, vol. 138, pp. 223–229, 2020.

[56] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[57] H. Li, P. Xiong, H. Fan, and J. Sun, “Dfanet: Deep feature aggregation for real-time semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9522–9531.

[58] H. Park, L. L. Sjösund, Y. Yoo, J. Bang, and N. Kwak, “ExtremeC3Net: Extreme lightweight portrait segmentation networks using advanced C3-modules,” 2019, *arXiv:1908.03093*.

[59] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and Chang Xu, “GhostNet: More features from cheap operations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1577–1586.

[60] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.

[61] Y. Yuan, X. Chen, X. Chen, and J. Wang, “Segmentation transformer: Object-contextual representations for semantic segmentation,” 2019, *arXiv:1909.11065*.

[62] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[63] H. Cao et al., “Swin-UNet: UNet-like pure transformer for medical image segmentation,” 2021, *arXiv:2105.05537*.

[64] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, *arXiv:2102.04306*.

Liguo Weng received the Ph.D. degree in electrical engineering from North Carolina A&T State University, Greensboro, NC, USA, in 2010.

He is currently an Associate Professor with the College of Automation, Nanjing University of Information Science and Technology, Nanjing, China. His main research interests include deep learning and its application in remote sensing image analysis.

Kai Pang received the B.S. degree in electrical engineering and automation and the Graduate degree majoring in electronic information from the Nanjing University of Information Science and Technology, Nanjing, China, in 2019 and 2022, respectively.

His research interests include deep learning and its application to remote sensing semantic segmentation.

Min Xia (Member, IEEE) received the Ph.D. degree in cybernetics control engineering from Donghua University, Shanghai, China, in 2009.

He is currently a Professor with the Nanjing University of Information Science and Technology, Nanjing, China. He is currently the Deputy Director of the Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing, China. His principal research interests include machine learning theory and its application.

Haifeng Lin received the Doctor’s degree in forest engineering from Nanjing Forestry University, Nanjing, China, in 2019.

He is currently a Professor with the College of Information Science and Technology, Nanjing Forestry University. His main research interests include networking, wireless communication, deep learning, pattern recognition, and Internet of Things.

Ming Qian received the Graduate degree in electronic information from the Nanjing University of Information Science and Technology, Nanjing, China, in 2021. He is currently working toward the Doctoral degree in computer science with Wuhan University, Wuhan, China.

His main research interests include deep learning and its application in remote sensing image analysis.

Changjie Zhu received the B.S. degree in philosophy from Hohai University, Nanjing, China, in 2008.

She is currently an Engineer with Hohai University. Her principal research interests include machine learning theory and its application.