SABNet: Self-Attention Bilateral Network for Land Cover Classification

Zhehao Hu, Yurong Qian, ZhengQing Xiao, Guangqi Yang, Hao Jiang, and Xiao Sun

***Abstract*—Land cover classification has been of great interest as one of the most prominent applications of remote sensing images. The emergence of convolutional neural networks has largely promoted the development of land cover classification, but it ignores the positional relationship between pixels. When remotely sensed features have both large intraclass scale differences and interclass similarities, it will result in the problems of fuzzy class boundaries of classification results and misclassification of small samples, which are difficult to be solved by existing methods. Inspired by the recent Transformer network, we propose a self-attentive bilateral network SABNet to alleviate these problems. Its backbone consists of a modified multiscale vision transformer and a stacked convolutional layer for extracting global spatial information and local contextual information. A local embedding module and a coordinate attention fusion module are further proposed in the feature fusion stage to reduce attention distraction and efficiently fuse the high and low features. A stepwise feature fusion module is proposed in the decoder to fully fuse the features extracted from the two branches. Experiments show that our method achieves the best results in mIoU on both Landcover.ai and GID-15 datasets with a similar number of parameters, 91.49% for the Landcover.ai dataset and 64.23% for the GID-15 dataset, compared with existing methods.**

***Index Terms*—Attention mechanism, bilateral network, land cover classification, remote sensing, transformer.**

I. Introduction

THANKS to the rapid development of aerospace technology, it is easier to acquire high-resolution remote sensing images with more detailed land surface information, which can support finer classification and increase the diversity of remote sensing applications, but how to effectively acquire the information contained in high-resolution remote sensing images has become an urgent problem to be solved. Land cover classification as a decoding method has gradually become an important task in high-resolution remote sensing image processing, and has wide applications in land-use planning [1], [2]; environmental monitoring and protection [3], [4]; land resource management [5], [6]; agricultural production management [7], [8]; and disaster monitoring [9], [10].

Nowadays, remote sensing image land cover classification is mainly divided into traditional machine learning methods and deep learning methods. The traditional machine learning methods are mainly based on image element and object classification methods, and the image element-based methods are mainly based on support vector machines [11], decision trees [12], as well as random forests [13], which classify each pixel in remote sensing images as a classification unit, but they only consider the spectral information of pixels, ignore the spatial relationship between pixels, and are vulnerable to noise disturbance. The object-based method [14] divides the regions with similar gray values of neighboring pixels into different scenes by some strategies, and then extracts and classifies attributes for each scene, and this method cannot achieve accurate results when object extraction is difficult. In contrast, deep learning methods slowly become the mainstream method for land cover classification of remote sensing images because of their strong contextual and spatial information extraction abilities and generalization abilities.

FCN [15] was the first fully convolutional neural network that was used for remote sensing image land cover classification, which performs end-to-end training but is unable to finely segment remote sensing images because its decoding part is too simple. UNet [16] uses a network structure of encoder-decoder with jump connections, which can better handle feature boundaries, however, for classes with large size differences in the scene. PSPNet [17], which utilizes pyramid pooling, has better results. After understanding the importance of global information for land cover classification, DeepLab [18] increases the perceptual field by using null convolution, and DANet [19] introduces a dual attention mechanism to capture remotely dependent semantic relations. However, these methods still rely on convolutional operations and cannot actually extract the global features.

Recently, the emergence of vision Transformer [20] has brought a new dynamism to the field of computer vision. Transformer was initially applied to natural language processing tasks, but in recent years it has been widely used in the image domain. Transformer can model the relationship between any two elements in a sequence by introducing a self-attentive mechanism [21] relationships to better capture global information,

Manuscript received 29 June 2023; revised 5 October 2023, 6 December 2023, 15 January 2024, and 3 March 2024; accepted 13 March 2024. Date of publication 27 March 2024; date of current version 24 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61966035, in part by the Major Special Project for High Resolution Earth Observation System of the National Defense Science and Industry Bureau under Grant 95-Y50G37-9001-22/23, and in part by the National Natural Science Foundation of China "Key Technology of Air-Sky-Earth Multimodal Data Fusion for Digital Twin Agriculture" under Grant 62266043. *(Corresponding author: Yurong Qian.)*

Zhehao Hu, Yurong Qian, Guangqi Yang, Hao Jiang, and Xiao Sun are with the School of Software, Xinjiang University, Urumqi 830008, China, and also with the Key Laboratory of signal detection and processing in Xinjiang Uygur Autonomous Region and Key Laboratory of Software Engineering, Urumqi 830008, China (e-mail: hzh867854083@163.com; qyr@xju.edu.cn; yanggqdlam@gmail.com; jh_2023@163.com; 107552104397@stu.xju.edu.cn).

ZhengQing Xiao is with the School of Mathematics and Systems Science, Xinjiang University, Urumqi 830008, China (e-mail: xiaozq@xju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3382096

Transformer has a more robust generalization capability than CNN and multilayer perceptron structures. However, the low induction bias and strong global receptive fields make it difficult for the Transformer model to adequately capture task-specific local details.

For these problems above, we propose a bilateral network based on a self-attentive mechanism for solving high-resolution remote sensing image land cover classification using BiseNet [22] as the benchmark network. It has two paths in the encoder part, one path is a spatial path that extracts global spatial information using the multiscale visual transformer ResTv2 [23], and the other path is a contextual path that consists of several stacked convolutional layers to extract local contextual information. To fully utilize the features extracted from these two paths, in the decoder part, we design a stepwise feature fusion module to fuse the features of the two branches. Considering that the shallow features of the Transformer still possess strong global information, the decoder with feature pyramid structure is used to fuse the features extracted from the encoder. To effectively delineate the boundaries and increase the generalizability of the model, we subsequently proposed the local embedding module and the feature fusion module based on coordinate attention [24] experimentally to reduce attention distraction and effectively fuse transformer high and low-level features. We demonstrate its effectiveness on the Landcover.ai dataset and its generalization on the GID-15 dataset. The key contributions of this article are as follows.

1) We pioneered a novel bilateral network consisting of convolutional and transformer blocks on Landcover.ai and GID-15 datasets, which can be effectively used for the land cover classification task, and demonstrated its effectiveness through comparative experiments.
2) We pioneered a local embedding module with global feature extraction constrained by convolutional blocks and a feature fusion module guided by coordinate attention for the Transformer network to reduce the distraction of the Transformer network.
3) We propose a step-by-step feature fusion module with a pyramidal feature structure that utilizes position and channel attention to efficiently fuse global and local features extracted from two paths and to fuse feature information at different scales.

The rest of this article is organized as follows. The related work is described in Section II. The proposed method is described in Section III. Section IV details the ablation and comparison experiments performed and their results. Finally, the summary is in Section V.

II. Related Work

A. Semantic Segmentation Network-Based Land Cover Classification

Semantic segmentation of remote sensing images is a hot problem in land cover classification, which improves the accuracy and efficiency of land cover classification by assigning each pixel in a remote sensing image to a defined class. This achieves automatic recognition and classification of different features in remote sensing images, and finely portrays the shape, texture, and spatial relationships of the features. FCN [15] is the first fully convolutional network for semantic segmentation, which enables image classification networks to be directly applied to semantic segmentation tasks at pixel level. Improved networks based on FCN, such as UNet [16], SegNet [25], and Deeplabv3+ [26], use an encoder–decoder structure combined with hopped connections to fuse feature information at different levels to improve segmentation accuracy, and PSPNet [17] and Deeplab [18] use multiscale information to improve segmentation accuracy and robustness.

Recent neural networks take the abovementioned networks as benchmarks and combine other modules to focus on solving some problems of remote sensing image segmentation. To address the problem that small features are difficult to recognize and detect, Liu et al. [27] designed a novel residual Atrous Spatial Pyramid Pooling structure, which reconstructs the null convolution using dilated attention convolution to obtain important multiscale semantic information and reduce the complexity of the network through residuals. To address the edge problem, Chen et al. [28] optimally segmented the boundaries by extending the multilevel feature aggregation network by adding a simple and effective dual-path feature refinement module before each upsampling module, which uses two independent branches to obtain features of different depths. To address the phenomenon of similarity between classes and differences within classes of remote sensing images, Wang et al. [29] proposed a multilevel feature fusion method for the fusion problem of multiple feature maps with size and semantic differences, using a method of up-sampling the input feature maps step by step and reweighting them according to the channels to reduce the impact of differences in semantic information. Ma et al. [30] introduced a fuzzy logic unit in a convolutional neural network to deal with the ambiguity and uncertainty of HRRS and introduced a conditional random field at the end to optimize the image segmentation results.

B. Attention Mechanisms in Image Semantic Segmentation Networks

Attention mechanisms have achieved good results in many visual tasks, such as image classification [31], target detection [32], and semantic segmentation [19]. In the attention mechanism, each input feature is given a weight that reflects how relevant the feature is to the task at hand [33]. By using the attention mechanism, the model can be made to focus more on the important regions in the image, thus improving the classification accuracy and detection precision of the model. Some of the more common attention mechanisms are self-attention [21], channel attention [34], and spatial attention [35]. Most of the following networks are combinations and improvements of the above attentions. Chen et al. [18] embed the squeeze-and-excitation attention module into the ASPP module by weighting the output of the ASPP module to further improve the accuracy of the model. Fu et al. [19] combined spatial attention to emphasize informative spatial regions and suppress irrelevant ones, as well as channel attention to capture the interdependencies between different feature channels and adjust their weights accordingly. Hou et al. [24] decomposed channel attention into two 1-D feature encoding processes that can capture remote dependencies along one spatial direction while retaining accurate location information along the other spatial direction.

For remote context dependencies, many attempts are beginning to be made, such as recurrent thrifty attention [36] and positional context aggregation [37] modules, to capture long-range dependencies efficiently and effectively. Kampffmeyer et al. [38] proposed a purely connected network ConnNet., which predicts the connection probability of each pixel to its neighbors by exploiting the multilevel cascading context and remote pixel relationships embedded in the image. Maire et al. [39] trained a convolutional neural network to directly predict pairwise relationships defining affinity matrices, and spectral embedding translates these predictions into globally consistent scene segmentation and character/ground organization.

Dosovitskiy et al. [20] segmented image data into small image blocks and then recombined these image blocks into sequences, which were then processed using self-attention, which was the first time transformer was used in computer vision, and VIT and subsequent work showed good results. Liu et al. [40] proposed a hierarchical transformer that uses shifted window computation to restrict self-attention to nonoverlapping local windows, while also allowing cross-window connections. The hierarchical structure that allows flexibility in modeling at different scales and has linear computational complexity relative to the image size. Xie et al. [41] proposed a novel hierarchical structure of the transformer encoder MIT for outputting multiscale features and aggregating global and local information from different layers with a simple MLP. He et al. [42] embedded the Swin transformer into UNet and encoded spatial information in Swin transformer blocks by establishing pixel-level correlations to enhance the feature representation of occluded objects. Wang et al. [43] proposed a dynamically scalable attention model combining convolutional and Transformer features, which can dynamically select the model depth according to the size of the input image, alleviating the problems of insufficient global information extraction in a single convolutional model as well as the computational overhead limitations of a pure Transformer model.

III. Methodology

In order to solve the problems of confusing classification of similar features and difficult recognition of small-scale features in land cover classification, we propose a semantic segmentation network SABNet based on the improvement of BiseNet. In this section, we first introduce the general structure of SABNet, and then introduce each part of the framework and the design ideas of each part in detail.

A. Model Overview

The overall network structure of SABNet is inspired by BiseNet and adopts a bilateral network structure, one for the spatial path and one for the contextual path. Considering the uniqueness of the Transformer, we adopt a jump connection akin to UNet in the decoder part. The overall structure of the model can be seen in Fig. 1. The backbone of the model consists of a stacked convolutional block and an efficient transformer network ResTv2 [23]. The ResTv2 network produces four feature maps of varying sizes, and the LE module further extracts features. The CFM module merges the features of the last two layers to generate three feature maps containing different information and sizes. The feature maps outputted from the stacked convolution block are further fused with the features fused by the CFM module in the first layer of the SFF module, where we use the module consisting of channel attention and spatial attention to optimize the feature information of the stacked convolution rapidly. Finally, the final segmentation results are obtained by upsampling and fusion operations with the SFF module.

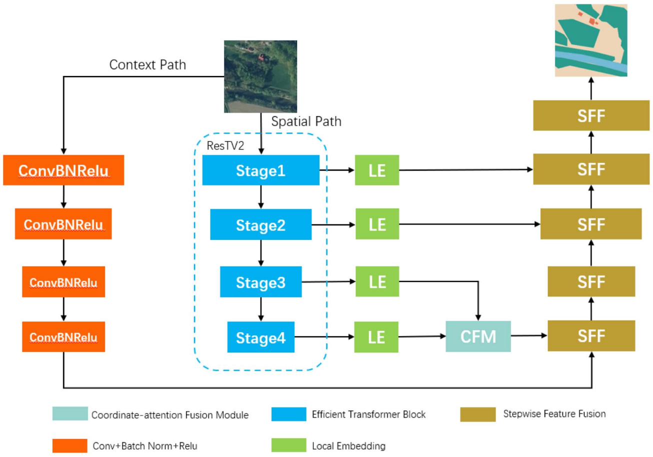

Fig. 1. General framework of SABNet.

B. Spatial Path

We use ResTv2-small as the backbone network in the spatial path, as an efficient visual transformer network that can efficiently capture long-range dependency information, balance the number of parameters and segmentation effects, and is well-suited for land cover classification of remote sensing images [23]. The structure of ResTv2 adopts a similar design idea to ResNet [44]. A stem module in the first layer extracts low-level feature information, followed by four stages to capture multiscale feature information. Each stage contains three parts: a patch embedding module, a location encoding module, and several efficient transformer modules. Specifically, at the beginning of each stage, the patch embedding module is used to reduce the resolution of the input token and expand the number of channels. The position encoding module is incorporated to suppress the positional information and enhance the feature extraction capability of the patch embedding. After completing these two stages, the input token is inputted into the efficient Transformer block.

The Stem module downsamples both the height and width of the input feature map by a factor of 4. To efficiently capture low-level feature information with fewer parameters, three convolutions are stacked separately using a convolution kernel of 3

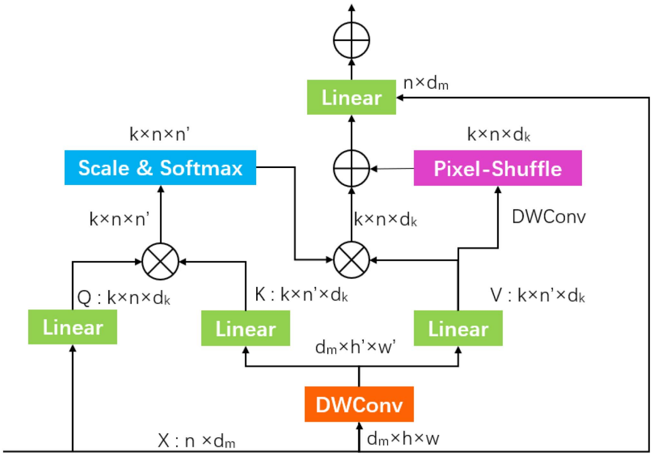

Fig. 2. Structural diagram of the EMSAv2 module.

and a padding of 1. The step size of [2, 1, 2] is added, and batch norm and ReLu are applied after the first two convolutions.

The patch embedding module downsamples the spatial dimension by 4 times, and the channel dimension becomes 2 times. This is achieved through a standard 3×3 convolution with a step size of 2 and padding of 1.

The position encoding module utilizes a more adaptable attention mechanism to acquire weights for individual pixels, and it applies a straightforward and effective pixel attention module for encoding positions. Precisely, pixel attention employs deep convolution to calculate pixel weights, and subsequently a sigmoid activation function for activation. The calculation is as follows:

$$\hat{x} = PA(x) = x \times \text{Sigmoid}(\text{DWConv}(x)). \qquad (1)$$

The efficient Transformer module uses the efficient multihead self-attention module EMSAv2, which is similar to the multihead self-attention module of transformer networks. EMSAv2 first uses a set of projections to obtain Q. To reduce memory, EMSAv2 reshapes the 2-dimensional input into a 3-D form before sending it to the depth convolution to reduce the k/8 The spatial dimensionality is then reduced by a factor of k/8, and the resulting features are reshaped into a 2-D form. They are then sent to the last two sets of projections to obtain K and V, after which the self-attentive calculation is performed. To solve the problem of losing too much information through EMSA downsampling operation, EMSAv2 then extends the channel dimension with a deep convolution for V and boosts the spatial dimension by pixel-shuffle [45] operation, which can effectively capture local information complementary to the long-range dependence and with fewer additional parameters and computational costs. Finally, the results of self-attention are summed with the up-sampling results to obtain the output of EMSAv2. EMSAv2 can be seen in Fig. 2, it is calculated as follows:

$$\text{EMSAv2}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V + Up(V). \qquad (2)$$

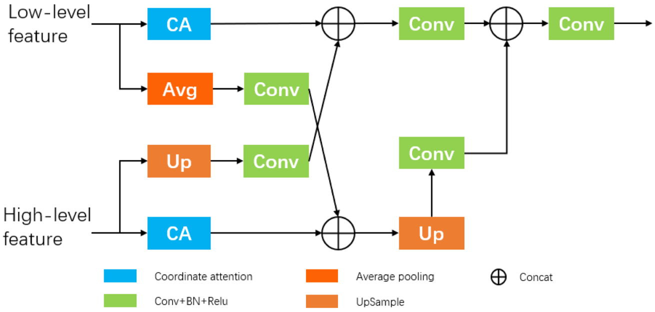

Fig. 3. CFM module structure diagram.

C. Context Path

The context path runs in parallel with the spatial way, extracting local contextual information missing from the spatial direction. As an auxiliary branch, we wanted to make it as light as possible, consisting of four convolutions

$$\hat{x} = CP(x) = \text{Conv}_4(\text{Conv}_3(\text{Conv}_2(\text{Conv}_1(x)))) \qquad (3)$$

where $\text{Conv}_i(\text{i} = 1, 2, 3, 4)$ all consist of Convolution, BN and Relu, the difference is that Conv_1 has a convolution kernel of 7, a stride of 2, a padding of 3 and the number of channels increases from 3 to 64. The other convolutional layers have convolution kernel of 3, a stride and padding of 2 and 1, respectively, and the number of channels is 64. Finally, the context path is downsampled by a factor of 8 and the number of channels is increased to 64.

D. Local Embedding Module and Coordinate Attention Fusion Module

It has been proposed that as the Transformer goes deeper, the attentional maps become increasingly similar and identical after specific layers. This suggests that in the deeper layers of the Vision Transformer model, the self-attentive mechanism cannot learn valid information for representation learning and prevents the model from achieving the expected performance, a phenomenon known as attentional collapse [46]. For this reason, we used the local embedding module LE, which utilizes the local perceptual field of the convolutional kernel to enhance the macroscopic weights around a block of query images, refocusing attention on neighboring features and thereby reducing attentional distraction. Specifically, the module consists of two 3×3 convolutions and Relu. The calculation formula is as follows:

$$\hat{x} = \text{Relu}(\text{Conv}(\text{Relu}(\text{Conv}(x))) + x. \qquad (4)$$

In addition, for better integration with contextual paths, we propose a Transformer high-level and low-level features fusion module CFM based on coordinate attention, as shown in Fig. 3. Coordinate attention can encode channel relationships and long-term dependencies with precise location information, enhancing the representation of learned features in the Transformer network. Specifically, high-level and low-level features are first passed through the coordinate attention module. Feature interaction is performed through upsampling and several 3×3

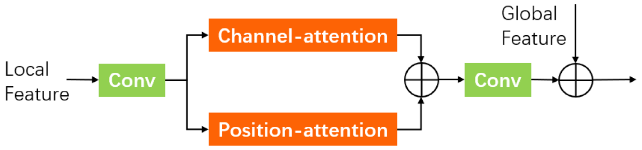

Fig. 4. SFF module first layer structure diagram.

standard convolutional blocks. Finally, the high-level features are upsampled and spliced with the low-level features to obtain the final result. The calculation formula is as follows:

$$X_H = \text{Conv}(\text{Avg}(X_{\text{LF}})) \oplus CA(X_{\text{HF}}) \tag{5}$$

$$X_L = (CA(X_{\text{LF}}) \oplus \text{Conv}(Up(X_{\text{HF}}))) \tag{6}$$

$$\hat{X} = \text{Conv}(\text{Conv}(Up(X_H)) \oplus \text{Conv}(X_L)) \tag{7}$$

where X_{LF} and X_{HF} denote low-level features and high-level features extracted by the Transformer, X_L and X_H denote features after the interaction of high-level and low-level features, respectively, CA represents coordinate attention, $\bigoplus$ symbols indicate stitching in the channel direction, Conv is 3 × 3 standard convolution (including BN and Relu). Avg and Up are 2× upsampling and 2× downsampling, respectively.

E. Stepwise Feature Fusion Module

Due to the weak correlation between features of different depths in the Transformer [47], layers of different depths need to interact with each other with a lot of information to guide each other. In order to merge global and local features extracted from both sides smoothly, we propose a decoder similar to a feature pyramid structure that fuses features progressively at different levels from bottom to top. It also enhances the representation of local features at the first layer by connecting it to the contextual path [48] using a double attention module consisting of parallel spatial attention and channel attention. Specifically, the SFF module has five layers, with the first layer using channel attention, spatial attention and two 3 × 3 standard convolution blocks, the structure of which can be seen in Fig. 4, and the remaining layers consisting of two 3 × 3 standard convolution blocks and one 2× upsampling block. The SFF module can be calculated as follows:

$$\hat{x} = \text{Conv}(CA(\text{Conv}(x)) \oplus PA(\text{Conv}(x))) \tag{8}$$

$$\hat{X}_i = Up(\text{Conv}(\text{Conv}(\hat{x}))) \oplus X_i \tag{9}$$

where CA and PA represent channel attention and spatial attention, respectively, $\bigoplus$ symbols denote splicing in the channel direction, x denotes features obtained after local features have passed through the first layer of the SFF module, $\hat{X}_i$ (i=1,2,3,4) denotes features after each layer has been fused, and X_i (docking with Transformer output features only when i is 2 and 3) denotes features after each layer of the Transformer network has passed through the LE module.

IV. Experimental Results and Analysis

A. Introduction to the Dataset

The Landcover.ai dataset [49] selected 41 orthophotos manually from different counties in various regions of Poland, each covering an area of about 5 km^2. In total, 33 images have a resolution of 25 cm (about 9000 × 9500 pixels) and eight have a resolution of 50 cm (about 4200 × 4700 pixels), with a total area of 216.27 km^2. These images are available in three bands of RGB. They are manually annotated for three categories: buildings, woodlands and water bodies, covering various rural areas and containing images of different optical conditions and vegetation seasons. In our experiments, we divided the dataset into a training set, a testset and a validation set according to the officially provided method of dividing the dataset. The divided dataset has 10 674 patches with a resolution of 512 × 512, of which 7470 patches are used for training, 1602 patches for validation and 1602 patches for testing.

GID-15 [50] subdivides the five land cover classification categories of GID-5 into the new GID-15. Built-up urban areas are subdivided into industrial, urban residential, rural residential, and transport land to study the detailed distribution of the urban regions. For the study of vegetation, forests are subdivided into parkland, tree and shrubland, and grasslands are subdivided into natural and artificial grasslands. Agricultural land is subdivided into paddy, irrigated, and dry land for the purpose of studying the distribution of agricultural land. GID-15 has the advantage of extensive coverage, wide distribution, good annotation, and high spatial resolution. The dataset comprises 10 RGB images and their corresponding annotations, each with a resolution of 6800 × 7200 pixels. We divided each image into nonoverlapping image blocks of size 512 × 512 pixels and discarded pixels at the edges that were not divisible by 512. In the end, 1820 patches were obtained, of which 1456 patches were randomly selected for training, 182 images for validation, and 182 patches for testing. Parkland, tree forest, shrubland, natural grassland, artificial grassland, paddy field, irrigated land, dryland, river, lake, pond, industrial land, urban residential, rural residential, and transport land.

B. Training Details

All our experiments were run on an NVIDIA RTX 3090 GPU and 24 GB RAM. The experimental code was implemented based on the PyTorch framework, where Python and PyTorch versions 3.8 and 1.11.0 were used. To optimize the network, we used AdamW as the optimizer, with the initial learning rate set to 0.0001. The minimum learning rate was also set to 0.0001, the momentum parameter was set to 0.9, the weights were decayed to 0.01, and the learning rate was decreased using cosine annealing. We performed random enhancement methods such as resizing, rotation, Gaussian blurring, and flipping on the input images. In terms of the loss function, we utilize the cross-entropy loss and the dice loss as loss functions to measure the difference between the segmentation results and the ground reference, where the dice loss corresponds to the global inspection while the BCE is a pixel-by-pixel micro-inspection,

TABLE I
IMPLICATIONS OF BCELOSS AND DICELOSS RATIOS

BCEloss:Diceloss	mIoU(%)	mPA(%)	mF1(%)
2	90.24	95.15	78.89
0.5	90.29	95.21	78.94
1	**91.50**	**95.46**	**79.50**

Bolding indicates the solution with the best result.

which can be complementary to each other. In Table I we experimented with the ratio of losses, and the optimal solution is a 1:1 ratio of the two losses. As the Transformer was used as the backbone network of the model, we considered using the pretraining weights of ResTv2-small on Imagenet-1 k to accelerate the model's training. Specifically, after loading the pretraining weights, we first freeze the backbone network and train the rest of the network, which has a batch size of 16. After freezing the weights for 50 epochs, the backbone network is unfrozen for uniform training; at this point, the batch size is 8. The formula for loss is as follows:

$$\text{Loss} = 0.5 \times \text{Dice loss} + 0.5 \times \text{BCE loss}. \qquad (10)$$

C. Evaluation Indicators

To quantitatively evaluate the accuracy of segmentation, we use three mainstream metrics: average F-1 score (mF1), average intersection/merger (mIoU) and average pixel accuracy (mPA) to evaluate the effectiveness of the network. MIoU is a region evaluation metric, and mPA and mF1 are pixel-level evaluation metrics. Parames refers to the number of parameters included in the model. FLOPS is a measure of processor performance and is an acronym for "floating point operations per second." These metrics are calculated on a cumulative confusion matrix with the equations shown as follows:

$$\text{mIoU} = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c}{TP_c + FP_c + FN_c} \qquad (11)$$

$$\text{mF1} = \frac{1}{N}\sum_{c=1}^{N}\frac{2 \times \frac{TP_c}{TP_c+FP_c} \times \frac{TP_c}{TP_c+FN_c}}{\frac{TP_c}{TP_c+FP_c} + \frac{TP_c}{TP_c+FN_c}} \qquad (12)$$

$$\text{mPA} = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c}{TP_c + FP_c} \qquad (13)$$

where TP_c is the number of positive samples correctly identified, FP_c is the number of negative samples misreported, TN_c is the number of negative samples correctly identified, and FN_c is the number of positive models missed. n is the number of categories in the dataset and c is a specific category.

D. Ablation Experiments

To validate the effectiveness of our proposed local embedding module, coordinate attention fusion module and stepwise feature fusion module; we conducted a series of ablation experiments based on the Landcover.ai dataset. In our experiments, we used BiseNet [22] as the base network and ResNet18 as the backbone network for spatial paths. In the ablation experiments, we improved the BiseNet network in turn to demonstrate the effectiveness of each module. We first changed the backbone network of BiseNet to ResTv2-small in the first experiment; then replaced the decoder part of BiseNet with a stepwise feature fusion module in the second experiment; then added the coordinate attention fusion module in the third experiment; and finally added the local embedding module as the fourth experiment, with the backbone network of the spatial path of each investigation for ResTv2-small. Based on the Landcover.ai dataset, different network variants were obtained by conducting experiments on BiseNet with the addition and replacement of network modules, and the results of evaluation metrics, such as mIoU, mPA, and IoU, were obtained, as given in Table II.

As can be seen from Table II, each of our improvements to BiseNet caused the evaluation metric mIoU to rise and reached a best mIoU of 91.50% and a best mPA of 95.46% by the time the upgrades were completed, fully demonstrating the effectiveness of our replacement network and the inclusion of modules. Specifically, when we replaced the backbone of the network from ResNet18 to ResTv2-small, the cross-merge ratio for each category increased and improved by approximately 2% each in the metrics mIoU and mPA, thus illustrating the effectiveness of the multiscale visual converter ResTv2 as a feature extractor. The stepwise feature fusion module, as a module for fusing features from both sides, was improved by 0.97% in mIoU using the feature pyramid structure and the dual attention module. The IoU values increased in each category, indicating the effectiveness of the decoder with the feature pyramid structure in feature fusion within the Transformer network. Moreover, the low-level features of the Transformer also play a role in this fusion. We replaced BiseNet's attention improvement module with the coordinate attention fusion module for better fusion of the Transformer's high-level and low-level elements, which allows for sufficient feature interaction through coordinate attention. Considering the distraction problem common to the Transformer model, we proposed a local embedding module to emphasise the local features of each image block; as a result, the IoU of vegetation slightly decreased, and the IoU of buildings increased, as vegetation is mostly continuous in large areas while structures are dense and independent, which shows that the local feature extraction ability of the model is enhanced and illustrates the effectiveness of the local embedding module. To better demonstrate the ablation experimental results, we visualized them. As shown in Fig. 5.

To demonstrate the effectiveness of the Transformer backbone we chose for the spatial paths, we compared it with the two most commonly used Transformer backbones, evaluated in terms of mIoU, Params, and Flops, and the results are given in Table III . It can be seen that ResTv2 is superior as a Transformer backbone.

E. CFM Module Fusion Experiment With Different Layers on Landcover.ai Dataset

Considering that the fusion of the CFM module in the last two layers is not the optimal result, in order to get its fusion effect in different layers, we designed this experiment, which are all the results obtained by running 100 epochs on the Landcover.ai

TABLE II
RESULTS OF ABLATION EXPERIMENTS ON THE LANDCOVER.AI DATASET

	backbone	SFF	CFM	LE	Background(%)	Water(%)	Woodlands(%)	Building(%)	mIoU(%)	mPA(%)
(a)	Resnet18				93.23	92.64	90.26	75.34	87.87	93.31
(b)	ResTv2-s				94.16	94.63	91.47	80.33	90.15	95.23
(c)	ResTv2-s	√			94.61	95.05	91.95	82.85	91.12	95.23
(d)	ResTv2-s	√	√		**94.79**	95.28	**92.21**	83.48	91.44	95.09
(e)	ResTv2-s	√	√	√	94.68	**95.30**	92.07	**83.93**	**91.50**	**95.46**

Bolding indicates the solution with the best result.

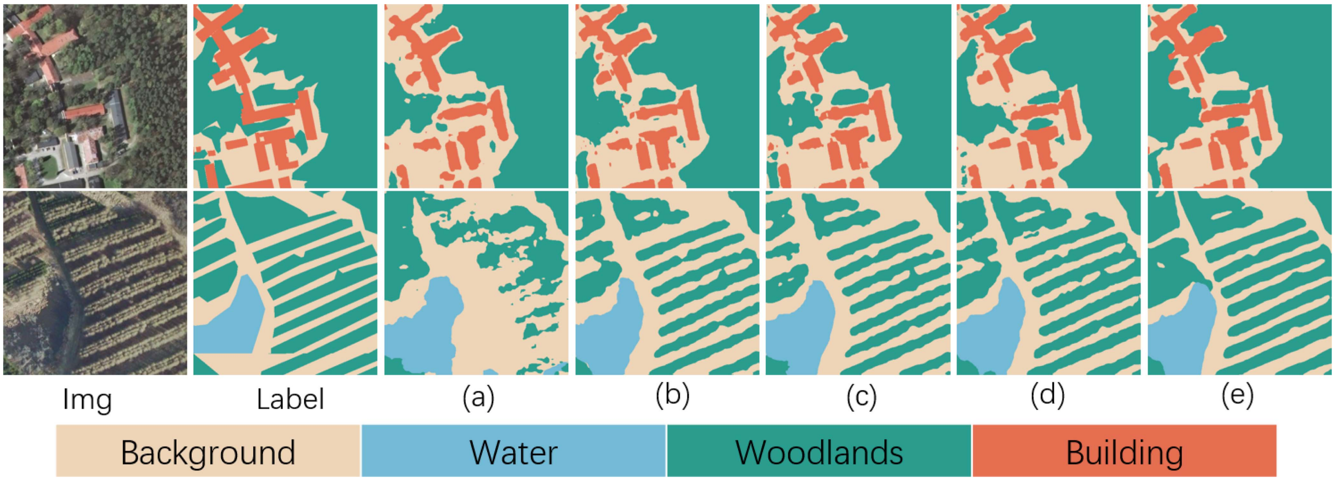

Fig. 5. Comparison of visualization results of different ablation models on the LandCover.ai dataset.

TABLE III
EXPERIMENTAL RESULTS ON SPATIAL PATHS OF DIFFERENT TRANSFORMER BACKBONE NETWORKS

Backbone	mIoU(%)	Params(M)	Flops(G)
Swin-s	88.61	63.82	73.76
MIT-b2	89.34	**41.41**	125.33
ResTV2-s	**91.50**	54.65	**70.47**

Bolding indicates the solution with the best result.

TABLE IV
RESULTS ON THE FUSION OF CFM MODULES AT DIFFERENT LAYERS

Layers	mIoU(%)	mPA(%)	Params(M)
1, 2	88.89	94.45	55.52
2, 3	90.69	95.03	56.13
3, 4	**91.50**	**95.46**	**54.65**

Bolding indicates the solution with the best result.

TABLE V
METRICS RESULTS IN THE LANDCOVER.AI DATASET COMPARED TO OTHER ADVANCED NETWORKS

Method	mIoU(%)	mPA(%)	mF1(%)	Params(M)	Flops(G)
UNet	89.30	93.60	94.96	43.93	92.02
SegNet	87.02	93.39	92.74	53.55	47.62
BiseNet	87.87	93.31	94.55	13.27	14.83
FCN-8s	82.39	89.53	91.80	134.28	191.28
PSPNet	86.95	93.20	94.47	49.07	61.47
DeeplabV3+	89.83	94.42	95.35	54.71	83.10
DIResUNet	75.22	87.30	85.97	20.64	492.39
Segformer	90.30	94.97	95.74	47.18	71.30
DEANet	90.28	94.54	94.81	60.29	-
HFENet	89.69	95.21	94.44	107.10	162.80
AMFFNet	90.39	94.66	93.94	66.13	52.10
SABNet	**91.50**	**95.54**	**96.14**	54.65	70.47

Bolding indicates the solution with the best result.

dataset. As we can see from Table IV the best results are achieved when fusing the last two layers of the backbone, which have a lower number of parameters as well as higher mIoU. This experiment shows that the advanced features of the transformer network have a better global feature extraction capability.

F. Comparison With Other Advanced Networks on the LandCover.ai Dataset

To demonstrate the advanced nature of our method, we conducted a series of comparison experiments on landcover.ai between SABNet and nine other state-of-the-art land cover classification methods. To demonstrate the fairness of the experiments, in addition to the benchmark network BiseNet, we selected network structures with similar parametric numbers as SABNet, namely: UNet [16], Deeplabv3+[26], PSPNet [17], SegNet [25], DIResUNet [51], FCN-8s [15], AMFFNet [48], HFENet [29], DEANet [52] and Segformer [41], and analyzed each network. UNet, DeepLabv3+, SegNet, DIResUNet and AMFFNet represent encoder–decoder networks. FCN-8 s represents full convolutional networks. PSPNet and DEANet represent networks with pyramidal ensemble methods. HFENet represents networks with attention mechanism methods. Segformer represents networks for the transformer method.

The results of the comparison experiments on the Landcover.ai dataset are given in Table V. We used three quantitative

TABLE VI
INDICATOR RESULTS FOR EACH FEATURE TYPE FOR COMPARISON EXPERIMENTS ON THE LANDCOVER.AI DATASET

Method	Background(%)	Water(%)	Woodlands(%)	Building(%)	Overall(%)
UNet	94.06	93.81	91.34	78.02	89.30
SegNet	93.54	93.67	90.90	69.99	87.02
BiseNet	93.23	92.64	90.26	75.34	87.87
FCN-8s	90.75	87.48	87.16	64.16	82.39
PSPNet	91.15	89.60	87.29	64.88	86.95
DeeplabV3+	93.29	91.48	88.91	73.98	89.83
DIResUNet	73.22	76.16	79.95	74.54	75.22
Segformer	94.34	94.61	91.64	80.60	90.30
DEANet	-	-	-	-	90.28
HFENet	94.28	94.19	91.62	81.97	89.69
AMFFNet	94.21	94.08	91.30	81.97	90.39
SABNet	**94.69**	**95.30**	**92.07**	**83.93**	**91.50**

Bolding indicates the solution with the best result.

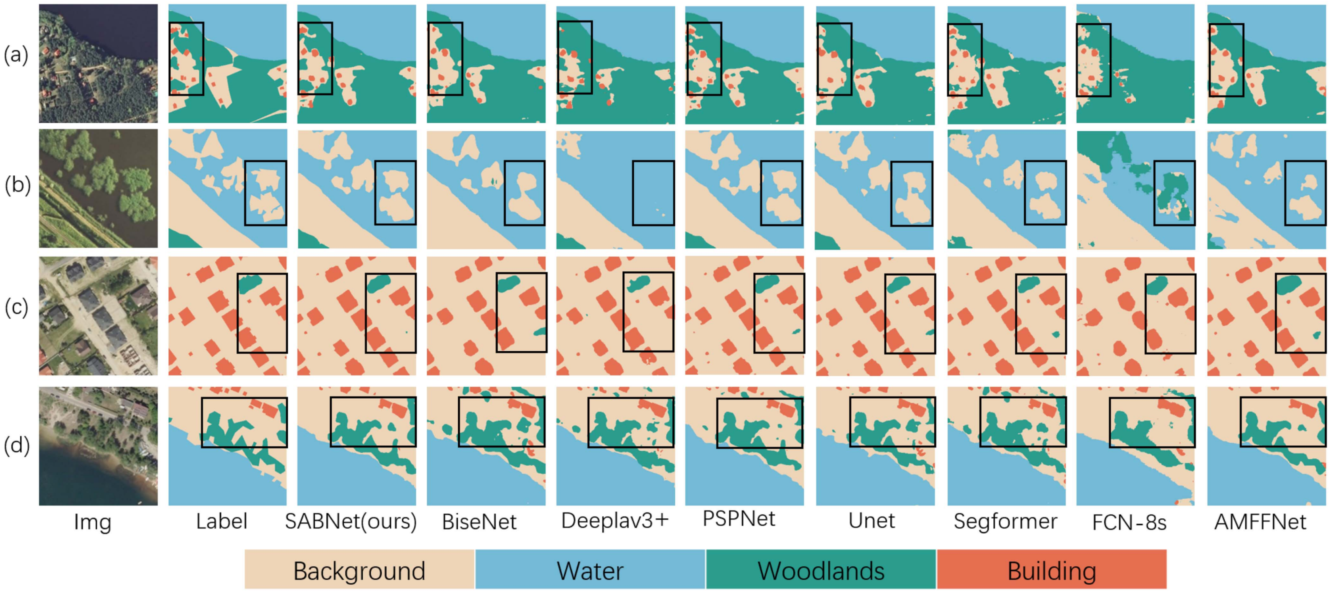

Fig. 6. Visual comparison of classification results from different advanced models on Landcover.ai.

metrics, mIoU, mPA and mF1, and Params and Flops of the model.

As can be seen in Table V, our proposed SABNet achieves the best results on the three evaluations with a lower number of parameters and computational effort. It is 1.11% higher in mIoU than AMFFNet, the best CNN network today, and 1.20% higher in mIoU compared with Segformer, the Transformer network used for segmentation. To better verify whether SABNet outperforms other methods in identifying features of complex and small remotely sensed features, we further counted the experimental results in each IoU value for each category, and the Params and Flops of the model, as given in Table VI.

In Table VI, it is easy to observe that SABNet achieves the best results for IoU in every category, especially for feature categories, such as buildings, which tend to occur on a small scale, with an improvement of 1.96%. For feature categories, such as water bodies and woodland, which are easily confused, there is also some improvement in the IoU values, illustrating the superiority of SABNet compared with other methods. To demonstrate the effectiveness of our method more visually, we visualized some of the model results. This is shown in Fig. 6.

It is known that for houses in wooded areas, it is easy to cause misclassification. SABNet can identify buildings well and delineate the boundaries of wooded areas better than other networks. When there are small pieces of land in the water, SABNet can extract the boundary information of the water body more accurately, while Deeplabv3+ and FCN-8 s cannot. When dense buildings were encountered for classification, other networks could not classify the buildings or poorly classified the boundaries. In contrast, SABNet could correctly identify small-scale targets, such as buildings with smooth boundary lines.

G. *Comparison With Other Advanced Networks on the GID-15 Dataset*

To further validate the generalization of the SABNet network, we conducted a series of comparison experiments using the GID-15 dataset with 15 classes and seven other advanced land cover

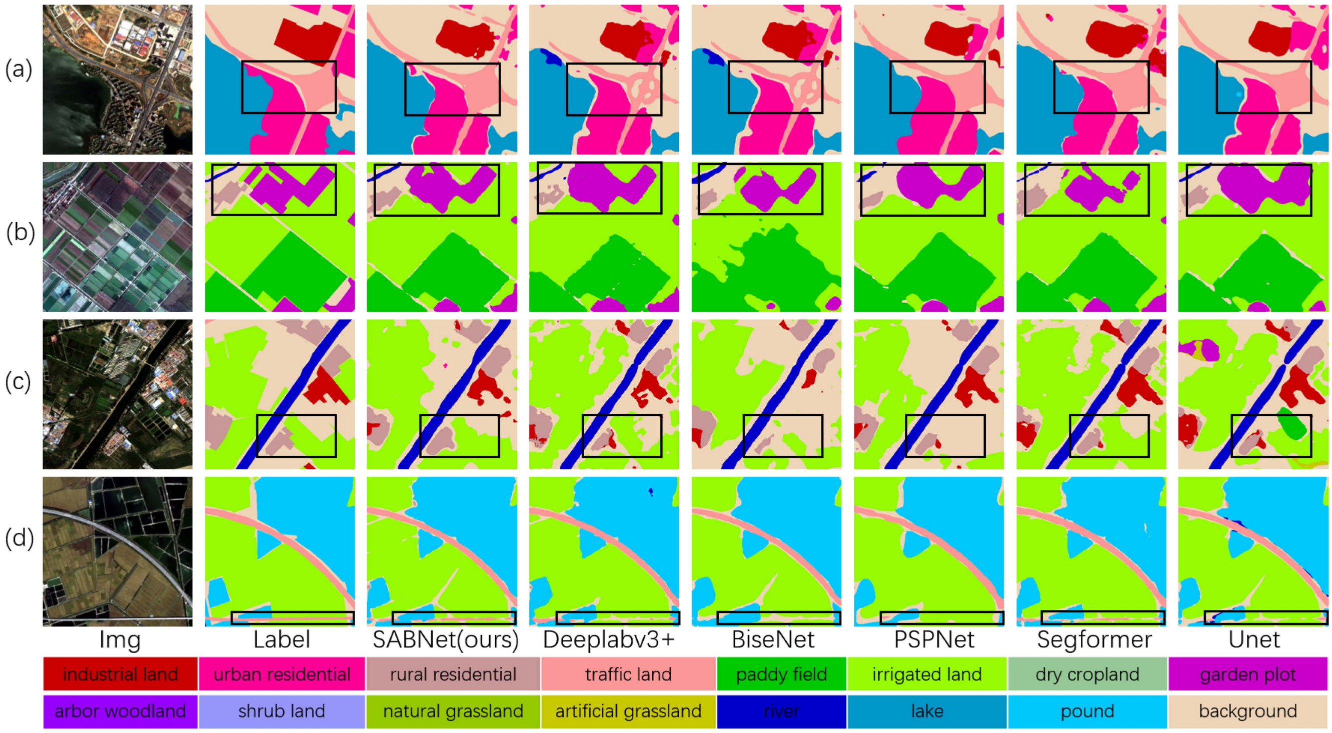

Fig. 7. Visual comparison of classification results of different advanced models on the GID-15.

TABLE VII
INDICATOR RESULTS FOR COMPARATIVE LEARNING IN THE GID-15 DATASET

Method	mIoU(%)	mPA(%)	mF1(%)	Params(M)	Flops(G)
BiseNet	56.07	69.60	74.30	14.88	13.27
PSPNet	60.35	69.64	75.47	49.07	61.47
Segformer	63.36	75.37	79.05	47.18	71.30
UNet	58.08	69.88	75.24	43.93	92.22
HPSNet	55.50	-	-	-	75.20
Deeplabv3+	60.30	75.95	78.48	54.71	83.15
AMFFNet	63.14	75.03	75.26	66.13	54.10
CAFHNet	62.50	-	-	8.50	-
SABNet	**64.23**	**76.49**	**79.50**	54.65	70.47

Bolding indicates the solution with the best result.

classification methods, namely, BiseNet [22], PSPNet [17], Segformer [41], UNet [16], Deeplabv3+ [26], AMFFNet [48], HPSNet [53], and CAFHNet [54], and calculated three quantitative metrics as well as the number of parameters and computation of the model for the different network experimental results.

As can be seen from Table VII, SABNet achieves the best results on the GID-15 dataset for the three quantitative metrics mIoU and mF1. With a similar number of parameters and computational effort, the primary metric mIoU surpasses today's best convolutional network AMFFNet by 1.09%, improves 8.16% over the benchmark network BiseNet, and increases 0.87% over the Transformer network Segformer, which fully illustrates the high generalization of SABNet. To illustrate more visually the superiority of SABNet over other methods, we visualized the classification results of some networks for comparison, and the results can be seen in Fig. 7.

In Fig. 7, we have chosen two classification maps to represent the unsmooth and error-prone classification of water body edges, and we can see that SABNet can extract water bodies and their boundaries better than other networks. In particular, the segmentation results of roads in this case also show that our network can identify finer roads more accurately. Furthermore, for similar feature classification and complex feature targets, SABNet can accurately segment the boundaries of similar features and determine the feature type.

V. CONCLUSION

In this article, we propose a novel and bilateral network for remote sensing feature classification based on the Transformer network, inspired by the BiseNet network's lightweight nature and the Transformer network's global nature. For issues related to the combination of Transformer and CNN networks, we propose new modules to enable better fusion of global and local features, such as a local embedding module and a coordinate attention fusion module, to reduce the distraction of the Transformer network and a stepwise feature fusion module to optimize the feature fusion process. Through the final ablation and comparison experiments, we found that SABNet can handle boundary information and similar feature information better, and the metrics also show that our network has fewer errors in feature classification.

REFERENCES

[1] Y. Tu, B. Chen, T. Zhang, and B. Xu, "Regional mapping of essential urban land use categories in China: A segmentation-based approach," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1058.

[2] Z. Fan et al., "Land cover classification of resources survey remote sensing images based on segmentation model," *IEEE Access*, vol. 10, pp. 56267–56281, 2022.

[3] B. A. Johnson and L. Ma, "Image segmentation and object-based image analysis for environmental monitoring: Recent areas of interest, researchers' views on the future priorities," 2020.

[4] Y. Cao and X. Huang, "A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 157–176, 2022.

[5] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[6] H. Lu, C. Liu, N. Li, X. Fu, and L. Li, "Optimal segmentation scale selection and evaluation of cultivated land objects based on high-resolution remote sensing images with spectral and texture features," *Environ. Sci. Pollut. Res.*, vol. 28, pp. 27067–27083, 2021.

[7] A. O. d. Albuquerque et al., "Deep semantic segmentation of center pivot irrigation systems from remotely sensed data," *Remote Sens.*, vol. 12, no. 13, 2020, Art. no. 2159.

[8] A. Sassu, F. Gambella, L. Ghiani, L. Mercenaro, M. Caria, and A. L. Pazzona, "Advances in unmanned aerial system remote sensing for precision viticulture," *Sensors*, vol. 21, no. 3, 2021, Art. no. 956.

[9] D. Q. Tran, M. Park, D. Jung, and S. Park, "Damage-map estimation using UAV images and deep learning algorithms for disaster management system," *Remote Sens.*, vol. 12, no. 24, 2020, Art. no. 4169.

[10] L. Yang, L. Wang, G. A. Abubakar, and J. Huang, "High-resolution rice mapping based on SNIC segmentation and multi-source remote sensing images," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1148.

[11] S. E. Jozdani, B. A. Johnson, and D. Chen, "Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification," *Remote Sens.*, vol. 11, no. 14, 2019, Art. no. 1713.

[12] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 140–152, 2020.

[13] L. Linhui, J. Weipeng, and W. Huihui, "Extracting the forest type from remote sensing images by random forest," *IEEE Sensors J.*, vol. 21, no. 16, pp. 17447–17454, Aug. 2021.

[14] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 115–134, 2019.

[15] 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.

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

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

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[19] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.

[20] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[21] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017.

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

[23] Q. Zhang and Y.-B. Yang, "Rest v2: Simpler, faster and stronger," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 36440–36452.

[24] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.

[25] 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.

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

[27] R. Liu et al., "RAANet: A residual ASPP with attention framework for semantic segmentation of high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3109.

[28] 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.

[29] D. Wang et al., "HFENet: Hierarchical feature extraction network for accurate landcover classification," *Remote Sens.*, vol. 14, no. 17, 2022, Art. no. 4244.

[30] X. Ma, J. Xu, Q. Chong, S. Ou, H. Xing, and M. Ni, "FCUNet: Refined remote sensing image segmentation method based on a fuzzy deep learning conditional random field network," *IET Image Process.*, vol. 17, no. 12, pp. 3616–3629, 2023.

[31] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3150–3158.

[32] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.

[33] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 347–356.

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

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

[36] L. Fu, D. Zhang, and Q. Ye, "Recurrent thrifty attention network for remote sensing scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8257–8268, Oct. 2021.

[37] D. Zhang, N. Li, and Q. Ye, "Positional context aggregation network for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 943–947, Jun. 2020.

[38] M. Kampffmeyer, N. Dong, X. Liang, Y. Zhang, and E. P. Xing, "ConnNet: A long-range relation-aware pixel-connectivity network for salient segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2518–2529, May 2019.

[39] M. Maire, T. Narihira, and S. X. Yu, "Affinity CNN: Learning pixel-centric pairwise relations for figure/ground embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 174–182.

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

[41] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[42] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[43] F. Wang, J. Ji, and Y. Wang, "DSViT: Dynamically scalable vision transformer for remote sensing image segmentation and classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 2023.

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

[45] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[46] D. Zhou et al., "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[47] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12116–12128.

[48] B. Tang, P. Tuerxun, R. Qi, G. Yang, and Y. Qian, "AMFFNet: Attention-guided multi-level feature fusion network for land cover classification of remote sensing images," *J. Appl. Remote Sens.*, vol. 17, no. 2, 2023, Art. no. 022205.

[49] 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.

[50] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.

[51] S. N. Priyanka, S. Lal, J. Nalini, C. S. Reddy, and F. Dell'Acqua, "DIResUNet: Architecture for multiclass semantic segmentation of high resolution remote sensing imagery data," *Appl. Intell.*, vol. 52, no. 13, pp. 15462–15482, 2022.

[52] H. Wei, X. Xu, N. Ou, X. Zhang, and Y. Dai, "DEANet: Dual encoder with attention network for semantic segmentation of remote sensing imagery," *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3900.

[53] K. Yang, X.-Y. Tong, G.-S. Xia, W. Shen, and L. Zhang, "Hidden path selection network for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[54] Y. Luo, J. Wang, X. Yang, Z. Yu, and Z. Tan, "Pixel representation augmented through cross-attention for high-resolution remote sensing imagery segmentation," *Remote Sens.*, vol. 14, no. 21, 2022, Art. no. 5415.

Zhehao Hu received the B.S. degree in big data analytics from the Hubei University of Technology, Wuhan, China, in 2021. He is currently working toward the master's degree in software engineering with Xinjiang University, Urumqi, China.

His research interests include remote sensing image segmentation and medical image segmentation.

Yurong Qian received the B.S. and M.S. degrees in computer science and technology from Xinjiang University, Urumqi, China, in 2002 and 2005, respectively, and the Ph.D. degree in Biology from Nanjing University, Nanjing, China, in 2010.

From 2012 to 2013, she was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Hanyang University, Seoul, South Korea. In 2015, she was trained as a Young Scientific and Technological Innovator with the Department of Science and Technology, Xinjiang, China. She is currently a Professor with the School of Computer Science and Technology and the School of Software, Xinjiang University. Her research interests include computational intelligence, such as Big Data processing, image processing, and artificial neural networks.

Dr. Qian is a Senior Member of the Chinese Computer Society.

ZhengQing Xiao received the Ph.D. degree in computer science and technology from Beijing Normal University, Beijing, China, in 2011.

He is currently working with the College of Mathematics and System Sciences, Xinjiang University, Urumqi, China. His research interests include big data analysis, image processing, and complex system modeling.

Guangqi Yang was born in Zhoukou, Henan, China in 1997. He received the Ph.D. degree in software engineering as a from the College of Computer Science and Technology, Jilin University, Changchun, China.

Since 2023, he has been interning with the State Key Laboratory of Automotive Safety and Energy and the School of Vehicle and Mobility, Tsinghua University, Beijing, China, and responsible for data visualization and article writing.

Hao Jiang received the B.S. degree in Internet of Things engineering from Henan Engineering College, Xinxiang, China, in 2020. He is currently working toward the master's degree in software engineering, Xinjiang University, Urumqi, China.

His research intersts include deep learning and spatio-temporal fusion of remote sensing images.

Xiao Sun is currently working toward the M.S. degree in software engineering with the School of Software, Xinjiang University, Urumqi, China.

His research interests include computer vision and remote sensing.