SPANet: Successive Pooling Attention Network for Semantic Segmentation of Remote Sensing Images

Le Sun, *Member, IEEE*, Shiwei Cheng, Yuhui Zheng, *Member, IEEE*, Zebin Wu, *Senior Member, IEEE*, and Jianwei Zhang

***Abstract*—In the convolutional neural network, the precise segmentation of small-scale objects and object boundaries in remote sensing images is a great challenge. As the model gets deeper, low-level features with geometric information and high-level features with semantic information cannot be obtained simultaneously. To alleviate this problem, a successive pooling attention network (SPANet) was proposed. The SPANet mainly consists of ResNet50 as the backbone, successive pooling attention module (SPAM), and feature fusion module (FFM). Specifically, the SPANet uses two parallel branches to extract high-level features by ResNet50 and low-level features by the first 11 layers of ResNet50. Then, both the high- and low-level features are fed to the SPAM, which is mainly composed of a successive pooling operator and a self-attention submodule, for further extracting deeper multiscale and salient features. In addition, the low- and high-level features after the SPAM are fused by the FFM to achieve the complementarity of spatial and geometric information. This fusion module alleviates the problem of the accurate segmentation of object edges. Finally, the high-level features and enhanced low-level features of the two branches are fused to obtain the final prediction results. Experiments show that the proposed SPANet achieves a good segmentation effect compared with other models on two remotely sensed datasets.**

***Index Terms*—Attention mechanism, convolutional neural network, remote sensing images, semantic segmentation, successive pooling.**

I. Introduction

SEMANTIC segmentation, which classifies each pixel into a category, is currently one of the research hotspots in the field of image processing, especially for remote sensing images. It now plays an irreplaceable role in many practical applications, such as natural disaster damage assessment [1], [2], precision agriculture [3], [4], urban planning [5], [6], and military reconnaissance [7], [8]. With the development of deep learning technology [9]–[11], a fully convolutional network (FCN) [12] was first applied to the semantic segmentation of remote sensing images. However, this kind of direct utilization of the neural network model [13], [14] in natural images to remote sensing images cannot achieve appealing results. The main reason is that compared with natural images, remote sensing images have their own characteristics: objects with fine structures in remote sensing images are either small in size or very slender in structure, such as cars and the edges of buildings, and shadows generated by buildings will also adversely affect them. Therefore, it is still a challenging task to accurately segment remote sensing images, especially for those small-scale objects and object edges.

Recently, a series of theoretically excellent research methods have been put forward for the semantic segmentation of remote sensing images. Some models are used for the segmentation of single category of objects. Irwansyah *et al.* [15] categorized buildings by improving the U-Net model, while Zhang *et al.* [16] combined U-Net and the attention mechanism to improve the accuracy of building segmentation. In a dense fusion classmate network [17], the lack of features in the network was compensated by joint training with a remote sensing dataset and a road dataset, and the accurate recognition of confused pixels was, thus, ensured. However, owing to the requirement of practical application, more models are dedicated to multicategory segmentation. Specifically, these models all exploit the continuity of contextual information, which ensures that the network models are extremely robust to objects at different scales in the dataset. In [18], features at different scales extracted through the pooling operation of the encoding stage and features at different scales extracted during the upsampling process of the decoding stage were fused to improve the segmentation accuracy of the road in the dataset. In dense dilated convolutions' merging network (DDCM-Net) [19], the network used dilated convolution to extract multiscale features and integrated local and global information to improve the model's ability to recognize objects with similar characteristics at different scales. In [20], a dense connection and FCNs used multiscale convolution kernels to increase the richness and diversity of extracted information, so that the features extracted by the network had stronger representative characteristics; this results in an improvement in the segmentation accuracy of semantic segmentation. In [21], the inception network structure was adopted, and the transposed convolution

Manuscript received February 9, 2022; revised March 16, 2022; accepted May 11, 2022. Date of publication May 16, 2022; date of current version May 30, 2022. This work was supported by the National Natural Science Foundation of China under Grant 61971233, Grant 62076137, Grant U20B2065 and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20211539. *(Corresponding author: Yuhui Zheng.)*

Le Sun is with the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: sunlecncom@163.com).

Shiwei Cheng and Yuhui Zheng are with the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: cs_chengsw@nuist.edu.cn; zhengyh@vip.126.com).

Zebin Wu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zebin.wu@gmail.com).

Jianwei Zhang is with the School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: zhangjw@nuist.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3175191

and dilated convolution were used to extract multiscale features to improve the performance of the network. In general, these models are based on extracting multiscale features to enhance the coherence of context information and the robustness of the network.

Although the extraction of multiscale features has a good effect on the segmentation accuracy, most network models obtain segmentation accuracy improvements at the cost of increasing the complexity of the model and introducing a large number of parameters; however, this often leads to excessive memory and time consumption during the training phase. To alleviate these problems, the attention mechanism was introduced to extract the main features in the image, suppress noise, and other useless information and also reduce the storage and training time. Therefore, many scholars have applied it to the field of semantic segmentation of high-resolution remote sensing images. For instance, in dual expectation–maximization attention network [22], the spatial expectation–maximization attention model simulated the interdependence of spatial features to obtain rich contextual information. Another module of the network enhanced the ability to express features through the interdependence between channels. In [23], a lightweight channel attention module used average pooling and maximum pooling to extract salient features to enhance the expressive ability of features. A local attention network (LANet) [24] used a one-time reduction in size and enlargement of features (using average pooling) to enhance the representation ability of small-scale object features. In the contextual transformer (CoT) network model [25], the authors proposed a CoT block based on self-attention, which can enhance the continuity of context information and extract salient features. Although those networks based on the attention mechanism can extract key salient features, they ignore the information contained in local small blocks in high-resolution remote sensing images.

Therefore, in this article, we propose a successive pooling attention network (SPANet) to simultaneously combine multiscale feature extraction and self-attention mechanism. Compared to the LANet, the SPANet utilizes a successive pooling operator and concatenates the intermediate pooling features at different scales to extract deeper semantic features. The information of intermediate features is retained to prevent excessive loss of information. The core idea of this successive pooling operation is to continuously zoom in on high-resolution remote sensing images as if holding a magnifying glass, more detailed and rich salient features at different scales will be exploited; moreover, the continuity of contextual information and the stability of the network model will be ensured. This successive pooling operation alleviates the problem that small-scale objects in remote sensing images are difficult to accurately subdivide. Besides, the high- and low-level features extracted by the backbone are subjected to in-depth feature extraction, and they are fused to achieve the complementarity of spatial and geometric information. This fusion approach alleviates the problem of accurate segmentation of object edges.

The main three contributions of the proposed SPANet can be summarized as follows.

1) In the successive pooling attention module (SPAM), an innovative successive pooling mechanism is proposed, which can obtain deeper features by effectively extracting and fusing multiscale features. This pooling method plays a very important role in the segmentation of small-scale objects in high-resolution remote sensing images and effectively improves the accuracy of semantic segmentation.
2) The SPAM proposed for semantic segmentation of high-resolution remote sensing images organically couples the attention mechanism with multiscale feature extraction. Among them, the attention mechanism can extract the salient features in the image while suppressing noise and useless information.
3) Using ResNet50 as a backbone, the SPANet model fully excavates deep and shallow features through two branches and effectively merges them to achieve the complementarity of spatial and geometric information. This feature fusion module (FFM) alleviates the problem of accurate segmentation of object edges. Experiments on the Potsdam and Vaihingen datasets verify the superiority of the SPANet over most other advanced semantic segmentation methods, especially in the segmentation of small-scale objects and boundaries.

The rest of this article is organized as follows. Section II introduces previous work related to semantic segmentation. Section III then describes the SPANet network model in detail. In Section IV, a detailed experimental evaluation and discussion of the SPANet is presented. Finally, Section V concludes this article.

II. Related Works

A. Encoder–Decoder Architecture

The encoder–decoder structure is widely used in various computer vision tasks [26]–[29]. The emergence of encoder–decoder structure in semantic segmentation was inspired by the FCN [12]. The specific purpose of the encoding stage was to extract the deeper semantic information of the image at the cost of reducing the image resolution. The decoding stage used an upsampling strategy to restore the low-resolution feature map to the original size and, finally, employed image reconstruction to output a predicted segmentation map the same size as the original image. In the decoding stage of the U-Net [30] model, the features of the image were introduced into the decoder using a jump connection, so that the decoder added the geometric features of the image in the process of image resolution restoration to recover the lost details of the image. In the SegNet model [31], the author recorded the location index in the process of sampling pooling and, then, upsampled the index positions recorded in the encoding stage to obtain sparse features. The convolution operation was then used to transform sparse features into dense features. The discriminative feature network model [32] integrated channel attention module and averaged pooling operation based on the encoder–decoder structure to enhance feature representation ability, so as to alleviate the problem of large differences within classes, based on extracting features using codec structure. The LANet [24] added other modules to enhance the ability of feature expression. The DDCM-Net [19] also used the encoder–decoder structure to extract the features and utilized the dilated convolution to enhance the feature representation.

It can be seen that the encoder–decoder structure is crucial in the process of feature extraction. Therefore, our model structure also uses the encoder–decoder structure to achieve preliminary feature extraction.

B. Multiscale Feature Extraction

The continuity of context information plays a key role in the segmentation of objects of different scales in scene semantic segmentation. The continuity of context information requires the network model to have robust recognition capabilities for objects with different structural scales, and at the same time, it can deal with the problems of large intraclass differences and small interclass differences in different scenes. For instance, the DeepLab [33]–[36] series of deep convolutional neural network models have achieved good results for semantic segmentation. The Deeplabv3+ [36] model extracted features by using dilated convolution, while in the pyramid scene parsing network (PSPNet) [37] model, the author used parallel pooling to extract features of different scales. These models have achieved good results for semantic segmentation of natural images. Inspired by the parallel pooling of different scales to extract features in [37], Yu *et al.* [38] proposed a model, which was composed of a convolution module and a pyramid pooling structure, to extract features at different scales so as to improve the model's ability to recognize different ground object categories. In [39], the training phase was divided into two stages. The first-stage network extracted deep-level semantic features from the original size image, and the second-stage network extracted low-level features from the cropped original image blocks and merged the features of different scales extracted into the two stages to enhance feature representation ability; different from the methods in [37] and [38], in this article, the original images are cropped at different scales and then input into the network for multiscale feature extraction. In addition, scale features are fused to improve the performance of the network. Cui *et al.* [40] proposed an adaptive multiscale feature learning module. This module was used to enhance the representation of weak boundary features, thus improving the segmentation accuracy of the model. These multiscale feature extraction operations are designed to enhance the coherence of context information, thereby improving the robustness of the network model. In those methods described above, multiscale feature extraction from high-resolution images is not thorough. There is still big room for improvement. Therefore, the proposed SPANet uses successive pooling operations to extract multiscale features. This idea of successive pooling is then embedded into the attention mechanism in the two-branch deep convolutional neural network model, so that more detailed and rich features can be extracted, and the ability of the model to recognize ground objects of different scales can be improved.

C. Attention Mechanism

Attention mechanism modules are added to models to emphasize the important information of the target object and suppress irrelevant details. Recently, an attention mechanism has been used for semantic segmentation of natural images and achieved good results. A squeeze-and-excitation network (SENet) [41] enhanced network performance by using only one average pooling to emphasize channel-level feature representation, which reduced interference from unrelated features. By using a single convolution kernel to extract a single feature, the selective kernel network [42] used convolutional kernels of different sizes to extract features in multiple branches and finally merged the output of each branch to produce richer features. There is no doubt that those methods increase the computational complexity of the models. To alleviate the problem, lightweight network architectures have been proposed, including convolutional block attention module (CBAM) [43] and bottleneck attention module (BAM) [44]. The former executed the channel attention module and the spatial attention module on the feature map in order and obtained the feature maps enhanced by the two modules. The CBAM then performed element-level multiplication using the original feature map and finally achieved the purpose of adaptive refinement. The BAM added the parallel branch results of the channel attention module and the spatial attention module to obtain an attention map that combines spatial and channel information, which improved performance when classifying natural images. The dual attention network (DANet) [45] was also an attention mechanism network divided into two branches. The two modules obtained long-range context information from the spatial dimension and the channel dimension, thereby enhancing the model's ability to express features. More recently, the attention mechanism has also been applied to the semantic segmentation of remote sensing images. Especially, in the attention-guided label refinement network [46], the channel attention mechanism was used to gradually refine feature maps of different scales to improve the accuracy of semantic segmentation of high-resolution remote sensing images. The gate module [47] simultaneously realized multiscale feature extraction and boundary restoration. The spatial relationship module and the channel relationship module proposed by Mou *et al.* [48] were used to learn the global relationship between feature maps, thus aggregating similar features to enhance the continuity of the context and the representation of different scale features. Compared to [48], the LANet [24] only implemented an average pooling operation to emphasize the partial key feature representations, which was obviously not enough for remote sensing images.

Building on the foundation of the research described above, the SPANet organically couples an attention mechanism with multiscale feature extraction. The former emphasizes key features, and the latter refines object boundaries to improve the segmentation results of our network model.

III. Proposed Method

In this section, we describe the SPANet architecture in detail. We will first provide a general overview of the network model and the core idea of dealing with semantic segmentation of remote sensing scenes and then describe the function of each module in detail.

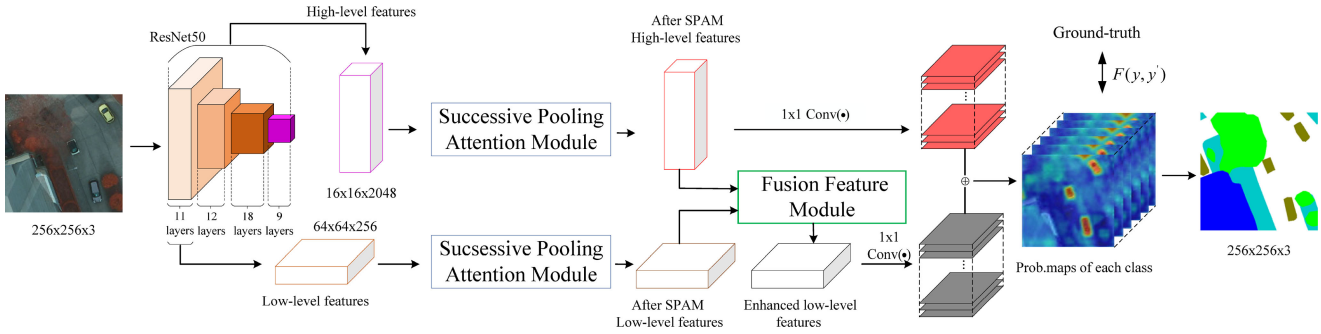

Fig. 1. Overview of the proposed SPANet architecture. The output of the 11th layer of ResNet50 is used as low-level features, and its size and number of channels are 64 × 64 and 256, respectively. The output of the 50th layer of ResNet50 is used as high-level features, and its size and number of channels are 16 × 16 and 2048, respectively.

A. *Overview of the Proposed SPANet*

The continuity of context information plays a key role in improving the accuracy of the semantic segmentation of remote sensing scenes. There is ambiguity about the features of different ground object categories in remote sensing images. Features extracted only by convolution and pooling operations cannot enable the network to recognize the features of different ground object categories; thus, the segmentation of object boundaries is easy to become blurred. Therefore, our proposed SPANet model uses a successive pooling method, which is much like browsing high-resolution remote sensing images with a magnifying glass. By analogy to the deep convolutional neural network model, we can extract more detailed and rich features that enhance the network's ability to recognize different object categories in remote sensing scenes. The architecture of the SPANet is shown in Fig. 1. The output of the 11th layer of ResNet50 is used as low-level features, and its size and number of channels are 64 × 64 and 256, respectively. The output of the 50th layer of ResNet50 is used as high-level features, and its size and number of channels are 16 × 16 and 2048, respectively. It is emphasized that we set the stride of the 43rd layer convolution operation in the original ResNet50 model from 2 to 1, and the purpose is to keep the size of the feature maps as 16 × 16.

The motivation for this article is as follows.

1) To use a method of successive pooling to realize the extraction of multiscale features. The detailed and rich features of high-resolution remote sensing images enhance the continuity of network context information and improve the network's ability to recognize similar feature categories.
2) Enriching the semantic representation of low-level features to better use spatial information.

These two points are reflected in the SPANet by the two modules: SPAM and FFM. The SPAM enhances the continuity of contextual information by extracting multiscale information in a modified attention module. The FFM fuses high-level semantic features and low-level features to realize the complementarity of spatial and geometric information. Specifically, we first use ResNet50 as our feature extractor, which is divided into two parallel branches after the feature extraction layer at different stages, called low- and high-level features. Immediately afterward, these are processed by the SPAM for the first enhancement of the features, and then, the high- and low-level semantic features after the first enhancement are used as the low-level features of the second enhancement after the FFM. Finally, the fusion of the first enhanced high-level semantic feature and the second enhanced low-level semantic feature is used as the final prediction segmentation graph. The detailed training framework of the SPANet model is shown in Algorithm 1.

B. *Successive Pooling Attention Module*

Unlike natural images, high-resolution remote sensing images have a large field of view and cannot show all the details of the object; in addition, some elements of the objects in the scene, such as low vegetation and trees, may be obscured because of the height and angle at which the image was captured. It is, therefore, likely to confuse human visual perception. For network models, actions should be taken to improve the model's ability to recognize objects with the same visual characteristics. In order to improve network performance, we use the SPAM to extract features at different scales in an attention module, thus improving the continuity of the context information, and to extract more detailed and salient features.

Fig. 2 shows a representation of the SPAM, which is inspired by the LANet [24]. We will next describe the specific operations of the SPAM. As shown in Fig. 2, the input 2-D feature map is assumed as $M \in R^{H\times W\times C}$. According to the structure of the transformer, the Keys, Queries, and Values are random matrices, and they are initialized as $Key = M$, $Query = M$, and $Value = M$. First, the feature matrix CK^1 is obtained by using the convolution kernel with kernel size $= 3$ and the convolution operation on the Key, and then, the feature matrix A is obtained by connecting CK^1 and Query after two 1×1 convolutions (KS_{1_1} with ReLU activation function and KS_{1_2} without activation function)

$$A = \left[CK^1,\ \text{Query}\ \right] KS_{1\_1} KS_{1\_2}. \tag{1}$$

For the Value branch, we first perform a successive adaptive average pooling operation on Value. Then, the Value of the feature map of multiple channels changes from the size of $H \times W$ to the feature matrix PF_i^C of the size of $PH_i \times PH_i$

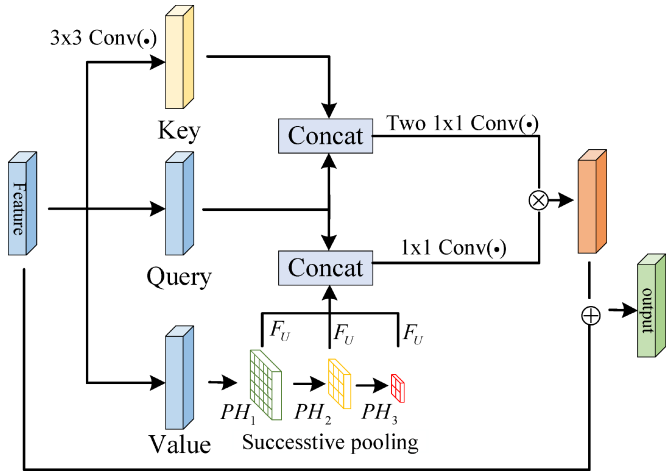

Fig. 2. Detailed design of the SPAM.

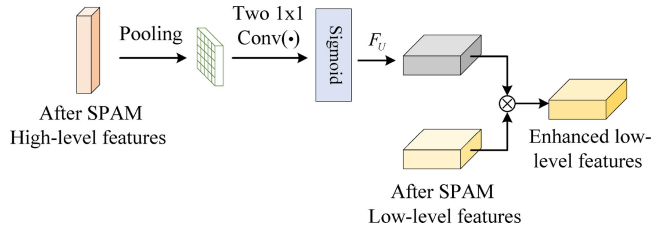

Fig. 3. Detailed design of the FFM.

after several successive pooling operations. The value PV_c of each grid point of each channel in the PF_i^C feature matrix corresponds to the average Value of the sum of all grid points in the size of $P_i \times P_i$ and the corresponding channel area in the original feature maps. This can be represented as follows:

$$P_i = H + 2 \cdot \text{ padding } - \text{ stride } (PH_i - 1). \tag{2}$$

According to the method of adaptive average pooling in the PyTorch library, stride $= 1$ and padding $= 0$. Therefore, the value of PV_c can be obtained by the following formula:

$$PV_c = \frac{1}{P_i P_i} \sum_{j=1}^{P_i} \sum_{k=1}^{P_i} x_c(j,k) \tag{3}$$

where x_c represents the value of a single grid point in the cth channel in the original feature map. With the calculation method of formula (3) and stride as 1, a feature matrix PF_i^C of the size of $PH_i \times PH_i$ can be obtained. In more detail, i is set to 1, 2, 3, and the corresponding PH_i is 10, 8, and 6, respectively; after the feature map is output by a pooling operation, the same pooling operation is performed twice in succession, and each time the output feature matrix is defined as PF_1, PF_2, PF_3. Next, an F_U performs operation on PF_1, PF_2, and PF_3 and the size becomes $H \times W$ (F_U is for upsampling operation). For the continuity of context information, we then concatenate the feature matrix of these three stages and Value and perform a 1 $\times$ 1 convolution operation to get PV. Next, we calculate the enhanced matrix CK^2 as follows:

$$CK^2 = PV \otimes A \tag{4}$$

where CK^2 is defined as the representative of the enhanced contextual information continuity of the input feature M. The symbol $\otimes$ represents channel element-level multiplication operation. According to the design of the residual block, it is conducive to the stability of the gradient backpropagation, and the final output enhanced feature matrix AF is the fusion of the original feature map M and CK^2:

$$AF = M \oplus CK^2 \tag{5}$$

where the symbol $\oplus$ represents channel element-level summing operation.

C. Feature Fusion Module

Blindly increasing the depth of the network and extracting deep semantic features can sometimes play a negative feedback role in the performance of semantic segmentation. Therefore, the high-level features containing spatial information should be superimposed on the low-level semantic features at the appropriate stage to achieve information complementation and improve semantic segmentation performance. In order to make full use of low-level semantic features, our proposed FFM adds spatial details from high-level semantic features to low-level semantic features, so that the model realizes the complementarity of spatial detail information and geometric information. Fig. 3 shows the details of the FFM. The FFM input is the high-level feature $AF_h \in R^{H' \times W' \times C_h}$ and the low-level feature $AF_l \in R^{H_l \times W_l \times C_l}$ after the SPAM. First, we perform an average pooling operation on the advanced feature AF_H as in formulas (2) and (3) to get feature matrix AP_h [PH_i is set to 10, padding $= 3$, and stride $= 1$ in (2)]. We then get the enhanced attention matrices AS_l by convolving the AP_h twice in succession

$$AS_l = F_U\left\{\left[KS_{1\_4}\left(KS_{1\_3}AP_h\right)\right]\right\} \tag{6}$$

where KS_{1_3} convolution is a buffer reduction of the number of high-level feature channels. In order to perform subsequent operations, KS_{1_4} convolution will change the number of channels in the middle feature matrix after KS_{1_3} operation so that it has the same size as AF_l ("()" and "[]" represent different numbers of 1 $\times$ 1 convolution operations, and "{}" represents the upsampling operation). Finally, the feature matrix AM_l is used as an enhanced low-level feature representative fused with high-level semantic information

$$AM_l = AS_l \otimes AF_l. \tag{7}$$

D. Fusion of the Outputs of Two Parallel Branches

After the FFM, the enhanced low-level feature representation, which contains rich semantic information, can be obtained. This greatly improves pixel-level classification tasks. The two branches of high- and low-level features maintain the same number of channels, dimensions, and sizes after the SPAM and the FFM. After the two branches, we use six 1 $\times$ 1 convolution kernels to complete the classification of each branch and sum the

Algorithm 1: Training Framework for the SPANet Model.

Input: Input a remote sensing image $I \in R^{H\times W\times C}$ and ground truth G.

Output: Predicted maps of the test dataset.

1: Set *batch_size* $= 5$ the weight attenuation of all learnable parameters is 2×10^{-5}, the maximum iteration number is $n = 10^8$, optimizer *Adam* (learning rate $= 8.5 \times 10^{-5}/\sqrt{2}$), Loss $= -\frac{1}{M}\sum_{k=1}^{M}\sum_{l=1}^{L} o_l^{(m)} \log(q_l^{(m)})E_l$ (10);
2: After data augmentation preprocessing, images with size of $256 \times 256 \times 3$ and their corresponding labels are obtained;
3: Train the SPANet network model;
4: **for** $i = 1$ to n **do**
5: Extract low-level features from the 11th layer of ResNet50, and extract high-level features from the last layer of ResNet50;
6: High-level and low-level features are, respectively, fed to SPAM to obtain AF_H, AF_L;
7: AF_H and AF_L are input to FFM together to get AM_L;
8: Fuse AM_L and AF_L to obtain the prediction results;
9: Calculate the loss between prediction results and labels, and update the parameters of the model;
10: Verify the performance of this weight;
11: When there is a higher mIou or mF1, save the weight.
12: **break**
13: Get the optimal training weight.
14: **end for**
15: Use test dataset with the trained model to get predicted maps.

feature maps of the corresponding channels in the two branches, as our final result.

IV. Experiments and Results

In this section, we elaborate on the experimental design and experimental results. This section includes descriptions of the two remote sensing datasets used in this study, image enhancement methods, and evaluation indicators. In order to verify the practicability of the modules of our model, an analysis of ablation experiments is also included. Finally, we compare the indicators and results with other methods, which shows that our method is capable of better segmentation results.

A. Datasets

We used two remote sensing scene datasets captured in Potsdam and Vaihingen. The first dataset is Potsdam, which consists of 38 tiles with a size of 6000×6000, with a ground resolution of 5 cm. Tiles are composed of red–green–blue infrared (RGB-IR) four-channel images. The dataset also includes a digital surface model (DSM) and a normalized DSM (nDSM). In this study, we only used IRRG data. The label data are divided into six categories: impervious surfaces, building, low vegetation, tree, car, and clutter/background. For evaluation, the 24 pictures were divided with labels into training set of 19 pictures, a validation set of two pictures, and a test set of three pictures.

The second dataset is Vaihingen, which consists of 33 tiles with an average size of 2100×2100, with a ground resolution of 9 cm. Tiles are composed of RGB-IR four-channel images. The dataset also includes a DSM and an nDSM. In this study, we only used IRRG data. The number of feature categories in the label data is the same as the number of categories in the Potsdam dataset. For evaluation, the 17 Vaihingen datasets containing labels were divided into a training set of 11 pictures, a validation set of two pictures, and a test set of four pictures.

B. Datasets Augmentation and Evaluation Methods

It is difficult to annotate the remote sensing dataset because different categories of ground objects present very complex visual effects, which is also an important reason for the small amount of annotated datasets. Therefore, the Potsdam and Vaihingen training sets employ random flip or mirror for data augmentation. In this study, we also used the albumentations library to enhance the data, and all the training images were normalized to $[0.0, 1.0]$ following data augmentation. We applied test time augmentation in the image inversion and mirroring stage. For these two datasets, we used a sliding window (with a size of 448 $\times$ 448 and a step size of 100 pixels) on the test data by averaging the prediction results of the overlapping area and stitching the results together as the final output.

The evaluation indicators provided by the publisher of the dataset are also used in some modeling methods. Mean IoU (mIou), mean F1 score, per-class F1 score, and overall accuracy (OA) are usually used as evaluation indicators for semantic segmentation. OA refers to the sum of the correct number of pixels predicted by all the categories divided by the total number of pixels, which reflects the total accuracy of the model's prediction. MIou is a standard measure of semantic segmentation, which shows the accuracy of network prediction for each category by calculating the average of the ratio of intersection and union of all categories. Especially for high-resolution remote sensing images, the level of the mIou index can better reflect the performance of network segmentation and network robustness. A certain type of F1 score is defined as the harmonic mean of accuracy and recall as follows:

$$\mathrm{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{8}$$

It is worth noting that the mean F1 score (mF1) and mean IoU (mIou) in these two datasets were calculated as the average metric for all classes except the clutter class.

C. Training Details

Adam was used as the optimizer to train the network, and except for bias, batch-norm parameter, and polynomial learning rate $(1 - (\text{cur\_iter/max\_iter}))^{0.9}$, the weight attenuation of all the learnable parameters was 2×10^{-5}, and the maximum iteration number was 10^8. On Potsdam and Vaihingen datasets, we used an initial learning rate of $(8.5 \times 10^{-5}/\sqrt{2})$, and we reduced the learning rate by 0.85 times after every 15 epochs. We applied

TABLE I
RESULTS OF DATA A DISPLAY OF THE PROPOSED SPANET ABLATION EXPERIMENT (ON THE POTSDAM DATASET)

	Module				Indicators		
	LLF	HLF	SPAM	FFM	mF1 (%)	OA (%)	mIou (%)
Ablation study model	√	√			89.90	88.60	81.71
	√	√	√		90.87	89.36	83.38
	√	√		√	90.55	89.28	82.85
SPANet	√	√	√	√	91.31	89.73	84.15

LLF stands for low-level semantic features; HLF stands for high-level semantic feature.

TABLE II
RESULTS OF DATA A DISPLAY OF THE PROPOSED SPANET ABLATION EXPERIMENT (ON THE VAIHINGEN DATASET)

	Module				Indicators		
	LLF	HLF	SPAM	FFM	mF1 (%)	OA (%)	mIou (%)
Ablation study model	√	√			88.94	89.69	80.33
	√	√	√		89.04	89.96	80.50
	√	√		√	88.96	89.95	80.37
SPANet	√	√	√	√	89.41	90.01	81.11

LLF stands for low-level semantic features; HLF stands for high-level semantic feature.

a cross-entropy loss function with median frequency balancing weights as defined in the following:

$$E_l = \frac{\text{median}\left(\{r_l \mid l \in L\}\right)}{r_l} \tag{9}$$

$$\text{Loss} = -\frac{1}{M}\sum_{k=1}^{M}\sum_{l=1}^{L} o_l^{(m)} \log\left(q_l^{(m)}\right) E_l \tag{10}$$

where E_l is the weight for class l, r_l is the pixel frequency of class l, $q_l^{(m)}$ is the probability of sample belonging to class l, and $o_l^{(m)}$ denotes the class label of sample m in class l. For these two datasets, our training set randomly sampled 5000 patches of 256 $\times$ 256 size from the original image as input and set the batch size to 5. It is worth noting that both our model and the other comparison models use the same data augmentation to augment the data.

D. Ablation Study

To validate the effectiveness of our model, we conducted ablation experiments on the two high-resolution remote sensing datasets. The ResNet50 was used as a backbone, and all the modules (including low-level feature, high-level feature, SPAM, and FFM) were conducted on the features extracted by the backbone. The first set of ablation experiments was the fusion of the high- and low-level features of the backbone. The second group of ablation experiments was to extract the high- and low-level features through the backbone and made them go through the SPAM, and the results of the two branches were fused as the prediction results. The third group of ablation experiments was to integrate the high- and low-level features obtained from the backbone into the FFM to obtain enhanced low-level features; then, the enhanced low- and high-level features extracted from the backbone were fused as the prediction results. The last set of experiments was SPANet.

The results of the ablation experiment on the Potsdam dataset are shown in Table I. The mF1 and mIou of the second group are 0.97% and 1.67% higher than those of the first group, respectively. This demonstrates that the performance of the network was improved by the adoption of the SPAM. The mF1 and the mIou of the third group were 0.65% and 1.14% higher than those of the first group, respectively. This shows that the high-level features make up for the missing spatial detail information in the low-level features, thereby improving the indicators. The last experiment is SPANet after adding both the SPAM and the FFM, we can see that mF1, mIou, and OA were 0.44%, 0.77%, and 0.37% higher than those in the second set of experiments, while 0.76%, 1.3%, and 0.45% higher than those in the third set of experiments.

The results of the ablation experiments on the Vaihingen dataset are shown in Table II. The comparison results of each group of experiments show the same trend as were found for the Potsdam dataset, and the indices of the second and third groups were still higher than those of the first group. After adding SPAM and FFM modules, the mF1, mIou, and OA of the SPANet were 0.37%, 0.61%, and 0.05% higher than those in the second group of experiments, while 0.45%, 0.74%, and 0.06% higher than those in the third set of experiments. The above ablation experiments validate that the use of the two modules together improves the experimental indicators more significantly.

To prove the effectiveness of our successive pooling operation, we also conducted ablation experiments on this module. Table III shows the results of these ablation experiments conducted on the Potsdam dataset. It can be seen that the mF1 of each category in the SPANet with no successive pooling decreased significantly, and mIou, mF1, and OA all decreased by 3.62%, 2.12%, and 1.9%, respectively. The results of this experiment as performed on the Vaihingen dataset follow the same trend as the previous dataset, as shown in Table IV. After removing the successive pooling modules, mIou, mF1, and OA decreased

TABLE III
DATA RESULTS BEFORE AND AFTER USING SUCCESSIVE POOLING (ON THE POTSDAM DATASET)

	Module					Indicators		
	Surface	Building	Low-veg	Tree	Car	mF1 (%)	OA (%)	mIou (%)
SPANet	91.61	94.90	87.09	88.85	94.08	91.31	89.73	84.15
No successive pooling	89.21	91.63	86.12	88.44	90.53	89.19	87.83	80.53

TABLE IV
DATA RESULTS BEFORE AND AFTER USING SUCCESSIVE POOLING (ON THE VAIHINGEN DATASET)

	Module					Indicators		
	Surface	Building	Low-veg	Tree	Car	mF1 (%)	OA (%)	mIou (%)
SPANet	92.16	94.87	82.79	89.15	88.11	89.41	90.01	81.11
No successive pooling	92.10	94.73	82.58	89.00	84.74	88.63	89.83	79.87

TABLE V
MIOU VALUES AS A FUNCTION OF THE SUCCESSIVE POOLING SCALES ON THE POTSDAM (THE SECOND LINE) AND VAIHINGEN (THE THIRD LINE) DATASETS

Successive pooling scales	(10,8,2)	(10,8,4)	(10,8,6)	(10,8,7)
mIou(%)	83.50	82.34	84.15	84.11
	80.19	80.46	81.11	80.78

by 1.24%, 0.78%, and 0.18%, respectively. It can be seen from Tables III and IV that the F1 score of all categories has been improved after successive pooling, especially the F1 score of the car has increased by 3.55% in the Vaihingen dataset and by 3.37% in the Potsdam dataset. It once again verifies that our SPAM can extract deeper semantic features while extracting salient and multiscale features, which is conducive to the segmentation of small-scale objects.

Through the comparison of Tables III and IV, it is easy to find that the increase in the indicators of the two datasets is different, which is mainly caused by the different resolutions of the two data. Specifically, for the Potsdam dataset, which has a higher resolution, all the indicators increased significantly with the addition of the successive pooling operation; this is a further indication that the successive pooling operation has a great contribution to the segmentation accuracy of high-resolution remote sensing images. Besides, after integrating successive pooling operations, the continuity of contextual information was strengthened, and the original prediction result map was improved.

In addition, to prove that the setting of successive pooling size is optimal in our experiment, we plot the mIou values as a function of the successive pooling size on the above two datasets in Table V, where the pooling sizes are (10, 8, 2), (10, 8, 4), (10, 8, 6), and (10, 8, 7). It can be seen from Table V that the successive pooling size used in our model with (10, 8, 6) is optimal, and the experimental results obtained are the best. As expected, different pooling sizes do affect the experimental results. Empirically, we recommend setting parameters of successive pooling as (10, 8, 6), which makes SPANet achieve appealing results in most high-resolution remote sensing images.

E. Qualitative Analysis of Features

In this subsection, we present a visualization of the partial segmentation results of the ablation experiment to verify the effectiveness of our proposed module. Fig. 4 shows the segmentation maps after adding the SPAM and the FFM in order. In the original backbone, there was no large receptive field for high-resolution remote sensing images, so the prediction of different features was very fuzzy. However, after the addition of the SPAM, the continuity of context information is enhanced, multiscale and salient features can be extracted, and the original predicted results are improved for small-scale objects. After the addition of the FFM module, the segmentation accuracy of object edge details in the predicted results has been improved. In general, our SPAM achieves the enhanced continuity of context information, and our FFM achieves the complementation of low- and high-level features, so these two modules have made a great contribution to the segmentation results.

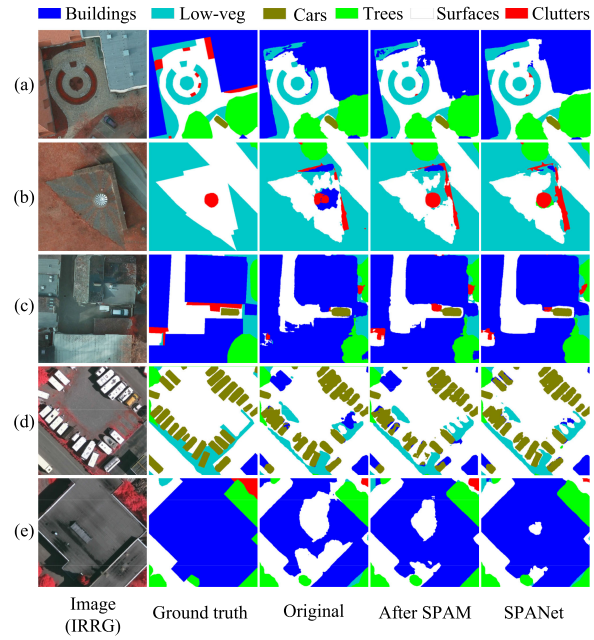

Fig. 4. After adding the SPAM and the FFM to the backbone in order, segmentation results of our network. (a)–(c) were cropped from the prediction results of the Potsdam dataset. (d) and (e) were cropped from the prediction results of the Vaihingen dataset.

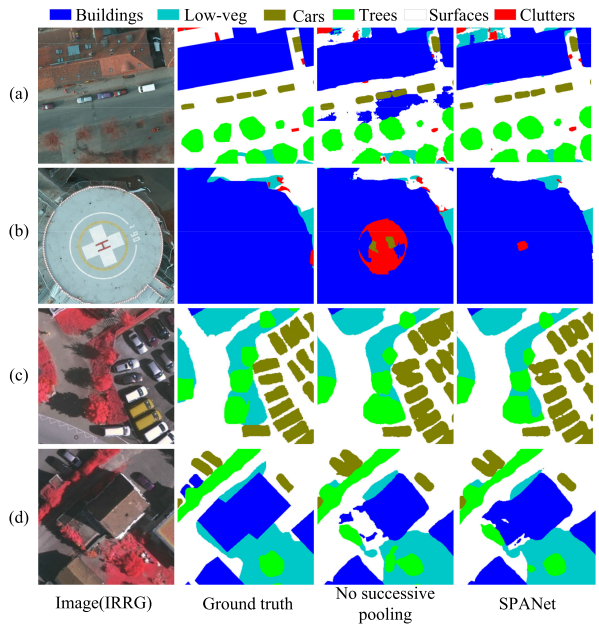

Fig. 5. Before and after adding successive pooling operation, segmentation results of our network. (a) and (b) were cropped from the prediction results of the Potsdam dataset. (c) and (d) were cropped from the prediction results of the Vaihingen dataset.

To prove the effectiveness of the successive pooling operation, we also performed an ablation experiment; the visual representation of this experiment is shown in Fig. 5. Compared

TABLE VI
DISPLAY OF VARIOUS DATA RESULTS OF DIFFERENT MODELS (ON THE POTSDAM DATASET)

Model	Per-class F1 score(%)					Indicators		
	Surface	Building	Low-veg	Tree	Car	mF1(%)	OA(%)	mIou(%)
CBAMNet [45]	88.60	90.11	81.47	82.97	87.04	86.04	85.14	75.64
Deeplabv3+ [39]	89.88	92.28	83.51	85.17	89.19	88.01	87.06	78.73
PSPNet [38]	90.55	91.88	83.31	85.75	89.31	88.16	87.24	78.97
SENet [41]	90.53	92.59	84.44	85.24	87.06	87.97	87.63	78.67
LANet [24]	90.58	93.04	86.48	88.73	93.10	90.39	88.87	82.56
DANet [45]	90.69	93.76	**87.59**	**88.96**	93.09	90.82	89.47	83.27
SPANet	**91.61**	**94.90**	87.09	88.85	**94.08**	**91.31**	**89.73**	**84.15**

It is worth noting that the bold font is the highest score.

TABLE VII
DISPLAY OF VARIOUS DATA RESULTS OF DIFFERENT MODELS (ON THE VAIHINGEN DATASET)

Model	Per-class F1 score(%)					Indicators		
	Surface	Building	Low-veg	Tree	Car	mF1(%)	OA(%)	mIou(%)
CBAMNet [45]	89.82	91.62	80.51	87.87	74.05	84.77	87.47	74.12
Deeplabv3+ [39]	90.62	92.51	79.18	87.26	79.25	85.77	87.71	75.50
PSPNet [38]	90.30	92.15	81.34	88.42	78.11	86.06	88.21	75.93
SENet [41]	90.65	92.55	81.53	88.60	79.70	86.61	88.46	76.73
LANet [24]	92.13	94.60	81.83	88.64	85.96	88.63	89.61	79.88
DANet [45]	92.13	**95.00**	82.48	88.94	85.96	88.90	89.91	80.30
SPANet	**92.16**	94.87	**82.79**	**89.15**	**88.11**	**89.41**	**90.01**	**81.11**

It is worth noting that the bold font is the highest score.

with the prediction maps without the successive pooling operation, these maps have improved a lot from the visual effect. Especially, the car category segmentation in rows (a) and (c) in Fig. 5 is clear. This operation also increases the continuity of context information through multiscale feature extraction and has strong recognition capabilities for different scale features in high-resolution remote sensing images, thus enhancing the robustness of the network.

F. Quantitative Comparison With State-of-the-Art Methods

We compared our proposed SPANet with other methods. In those methods, the backbone used to extract high- and low-level features was also ResNet50. The models contrasted with this experiment were derived from the most recent work on the attention mechanism, including CBAMNet [43], SENet [41], LANet [24], and DANet [45]. The experimental data results are shown in Tables VI and VII. The PSPNet [37] and DeeplabV3plus [36] models improve the accuracy of semantic segmentation by increasing the receptive field, which is also included in the model of our comparison experiment. Tables VI and VII show the results of different models applied to the Potsdam and Vaihingen datasets. From the results, we can see that for the Potsdam dataset, the Deeplabv3plus and PSPNet models achieved higher mF1 and mIou than the SENet and CBAMNet models. The two former models are based on the extraction of multiscale information. In the DeepLabv3plus model, spatial pyramid pooling and empty convolution are used to extract multiscale features, so as to capture clearer target boundaries by gradually recovering spatial information. PSPNet adopts parallel and average pooling methods to extract multiscale features, which enhances the continuity of context information and strengthens the model's ability to recognize objects at different scales. Although these two classic models have good semantic segmentation effects when applied to natural scenes, the effects are not ideal when the models are applied to high-resolution remote sensing images. The reason is that a parallel feature extraction is not thorough enough for extracting features from high-resolution images. The rich and detailed features are not extracted, so the segmentation of the boundary details of the image is not accurate. However, when applied to the Vaihingen dataset, the mF1, OA, and mIou of SENet are higher than those of PSPNet and Deeplabv3plus. This is because the Vaihingen dataset has a smaller resolution than the Potsdam dataset. The SE block using only one pooling operation achieved good results.

The patch attention module in the LANet uses an average pooling operation to make local features more prominent, thereby increasing the continuity of context information. However, only using one pooling operation for extracting deeper features from high-resolution remote sensing images is not thorough. Therefore, three successive pooling operations are used in the SPANet to extract deeper semantic features. Not only are the mF1, OA, and mIou in the SPANet (as applied to the Potsdam dataset) 0.92%, 0.86%, and 1.59% higher than those of the LANet, but the mF1, OA, and mIou in the SPANet as applied to the Vaihingen dataset are 0.78%, 0.4%, and 1.23%

TABLE VIII
COMPARISON RESULTS OF COMPLEXITY, NUMBER OF PARAMETERS, AND INFERENCE TIME (MEASURED ON INPUT IMAGE SIZE OF $3 \times 256 \times 256$) FOR DIFFERENT MODELS

Model	CBAMNet [45]	Deeplabv3+ [39]	PSPNet [38]	SENet [41]	LANet [24]	DANet [45]	SPANet
Params(Mb)	24.48	40.41	59.71	26.30	23.79	47.44	24.03
FLOPs(GFLOPS)	8.3	11.69	118.31	5.48	5.47	6.91	8.54
Inference time(ms-CPU/GPU)	937 / 53.99	1111 / 48.10	7809 / 180.20	671 / 62.07	577 / 59.86	728 / 70.11	865 / 55.72

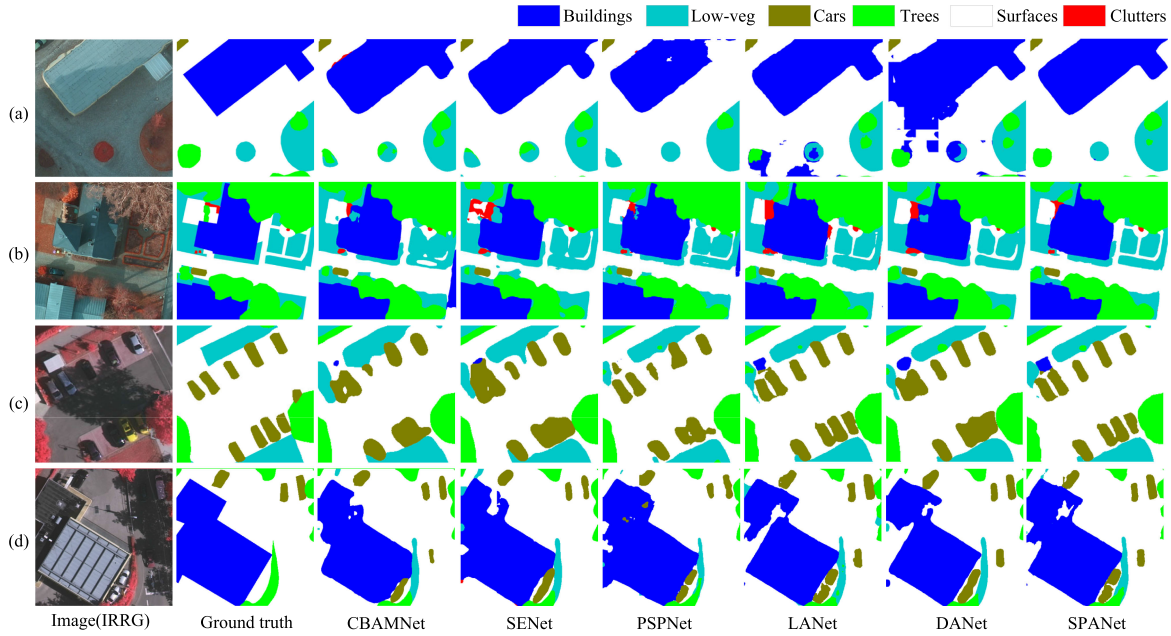

Fig. 6. Segmentation maps of all competing methods. (a) and (b) were cropped from the prediction results of the Potsdam dataset. (c) and (d) were cropped from the prediction results of the Vaihingen dataset.

higher, respectively. The DANet contains a position attention module and a channel attention module. The function of the former is to associate similar features of the same channel, and the function of the latter is to integrate the correlation of features between channels so as to strengthen the representation of features and obtain more accurate segmentation results. The DANet model applied to the Potsdam dataset is 0.5% and 0.11% higher for low vegetation and tree of F1 score, respectively, than the SPANet and 0.13% higher for building of F1 score than our model as applied to the Vaihingen dataset. From Tables VI and VII, it can be seen that not only our model is higher than other methods in the comprehensive indicators, but also the F1 score of each category is almost the highest. In addition, the F1 score of car in the Vaihingen dataset is 2.15% higher than that of the DANet. The improvement is significant. These phenomena show that the SPANet is better than the DANet in the segmentation of small-scale objects. In general, the proposed SPANet model performed better on the Potsdam and Vaihingen datasets than other methods, as measured in terms of mF1, OA, and mIou.

We compared the floating-point operations (FLOPs) and parameters required by the SPANet and other models. Table VIII shows the values of these two indicators. This calculation was based on a three-channel image with an input size of 256×256. The results clearly show that the attention-mechanism-based models CBAMNet, SENet, and LANet are lightweight models. DeepLabv3+ and PSPNet, on the other hand, require more computing resources but perform poorly on high-resolution remote sensing images. The DANet introduces a large number of learnable parameters to improve the performance of the model, but the number of parameters for the SPANet is much lower, and only a small amount of calculation is added in exchange for the improved segmentation performance of the network.

We also tested the time consumption of forward propagation of different models. The CPU is Intel (R) Xeon (R) Silver 4210, and the GPU is NVIDIA Geforce GTX 2080Ti. As shown in inference time in Table VIII, the more complex the model, the longer time required for forward propagation. Compared with the LANet and the DANet, the SPANet greatly improves the segmentation accuracy while reducing forward propagation time on the GPU.

G. Quantitative Analysis of Visualization Results of Different Models

Fig. 6 shows a comparison of the enlarged results of the SPANet and other methods. It is clear that models based on attention mechanisms, such as CBAMNet, SENet, and DANet, are not ideal for the semantic segmentation of high-resolution remote sensing images, especially for small objects and object

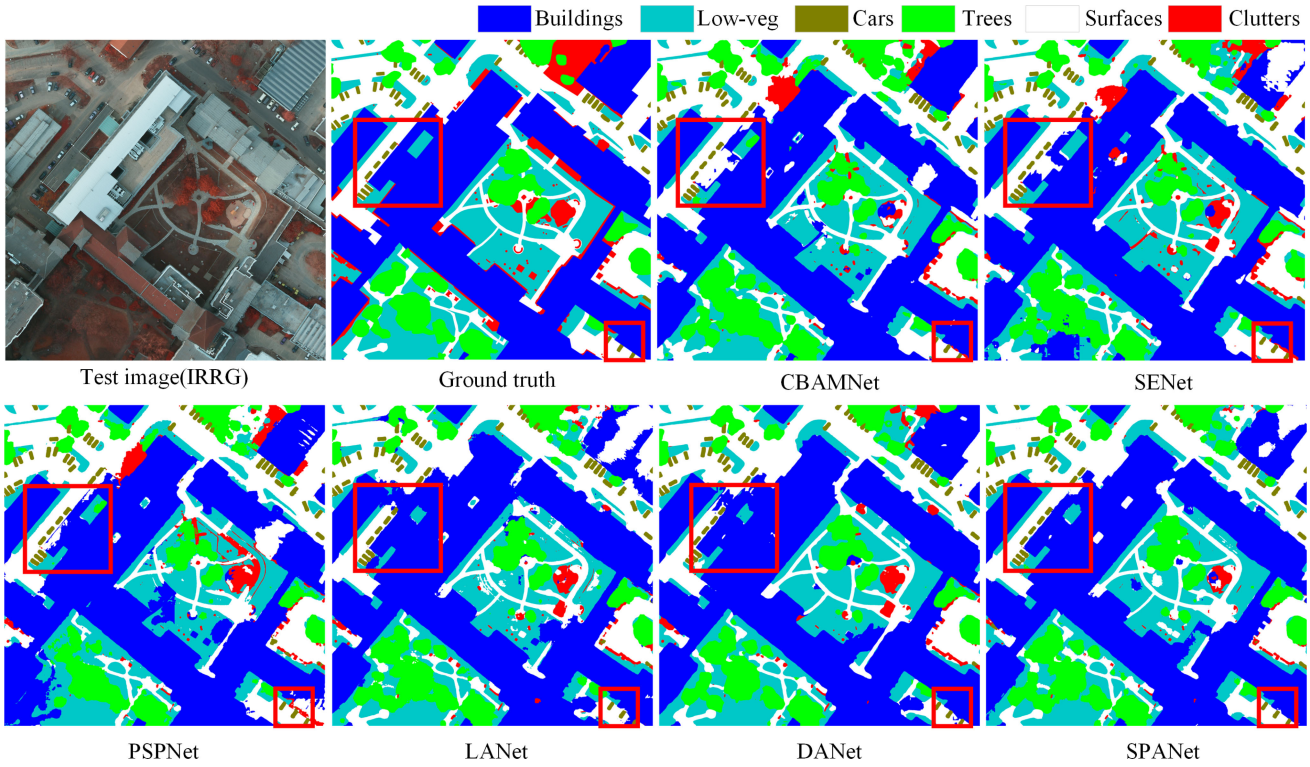

Fig. 7. Segmentation maps of all competing methods. These maps were cropped from the prediction results of the Potsdam dataset.

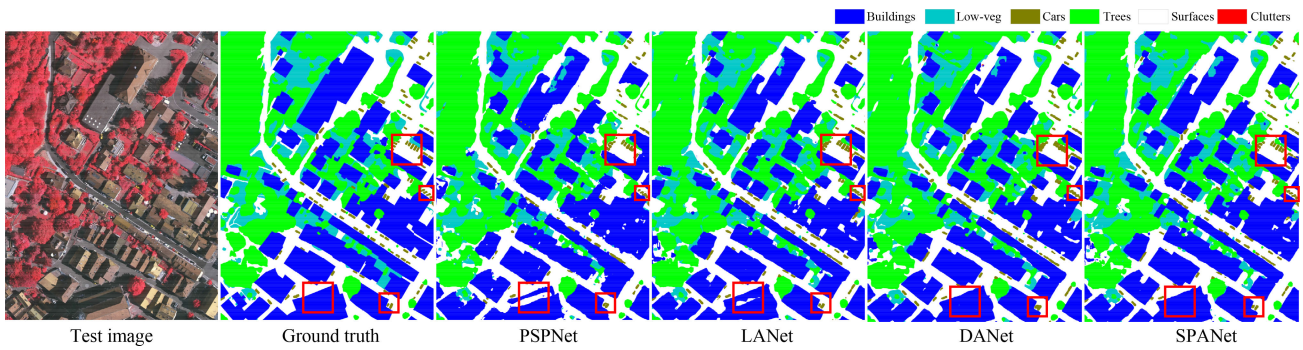

Fig. 8. Segmentation maps of all competing methods. These maps were cropped from the prediction results of the Vaihingen dataset.

boundaries. In the PSPNet, although the multiscale feature extraction method is used to increase the continuity of context information, it can be seen from the segmentation maps that there are misclassifications between different categories. In order to enhance the continuity of context information and extract more detailed and salient features, the proposed SPANet model achieves the purpose of complementing spatial and geometric information. Thus, object boundary segmentation is more accurate than other models. Moreover, the successive pooling operation is also very beneficial to the segmentation of small-scale objects; it can be seen from the comparison of the output results of different models in rows (a), (b), and (c) in Fig. 6. In addition, the segmentation effect of the SPANet for buildings, low-veg, and tree categories is also improved in comparison with other models.

Figs. 7 and 8 show large-scale prediction maps for the Potsdam and Vaihingen datasets. Although the DANet produces segmentation results that are better than those produced by other methods, it is still ambiguous for small-sized objects and boundary parts. In the result maps of SPANet semantic segmentation, the classification of errors between classes is very small; this is due to the fusion of high- and low-level semantic features. Meanwhile, the segmentation of object boundaries is very fine, which is due to the deep multiscale feature extraction in the SPAM. In general, the SPANet performs well in the segmentation of small-scale objects and object edge details.

V. Conclusion

In this article, we combined the attention mechanism with deep multiscale feature extraction, so the network can not only extract the key information from the image but also refine the details of objects. Specifically, the SPAM is proposed to enhance the representation of high- and low-level semantic features, and the FFM also makes spatial and geometric information complementary to each other.

1) The SPAM uses the attention mechanism to enhance the extraction of useful information to suppress noise and useless information, which is particularly important for remote sensing images of complex scenes. The successive pooling operation was embedded into the SPAM to extract multiscale and salient features, just like holding a magnifying glass and continuously zooming in and browsing across high-resolution remote sensing images.
2) The FFM alleviates the bottleneck problem in semantic segmentation. By fusing the semantic information in the deep network into the features of the shallow network, the features of the shallow network contain not only geometric information but also rich spatial information. The fusion of high- and low-level features makes the boundaries of objects more refined.

The results of the proposed SPANet on Potsdam and Vaihingen datasets show that the features extracted from the backbone were extracted into multiscale features after the SPAM so that a stronger feature representation was obtained. Therefore, compared to other models, good segmentation results were achieved, especially for small-scale objects and boundaries. However, the segmentation results for those parts affected by shadows in the scene are unsatisfactory. To conquer this problem, we plan to embed feature enhancement strategies to the SPAM for improving the representation ability of high-level features in the network.

References

[1] T. Chowdhury, M. Rahnemoonfar, R. Murphy, and O. Fernandes, "Comprehensive semantic segmentation on high resolution UAV imagery for natural disaster damage assessment," in *Proc. IEEE Int. Conf. Big Data*, 2020, pp. 3904–3913.

[2] Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, and T. Yin, "L1-norm distance linear discriminant analysis based on an effective iterative algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 114–129, Jan. 2018.

[3] T. Anand, S. Sinha, M. Mandal, V. Chamola, and F. R. Yu, "AgriSegNet: Deep aerial semantic segmentation framework for IoT-assisted precision agriculture," *IEEE Sens. J.*, vol. 21, no. 16, pp. 17581–17590, Aug. 2021.

[4] Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, and G. Yang, "Nonpeaked discriminant analysis for data representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3818–3832, Dec. 2019.

[5] Z. Guo *et al.*, "Semantic segmentation for urban planning maps based on U-Net," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6187–6190.

[6] Q. Ye *et al.*, "L1-norm distance minimization-based fast robust twin support vector k-plane clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4494–4503, Sep. 2018.

[7] J. Li, S. Gou, R. Li, J.-W. Chen, and X. Sun, "Ship segmentation via encoder-decoder network with global attention in high-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art no. 4016605.

[8] L. Sun *et al.*, "Low rank component induced spatial-spectral kernel method for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3829–3842, Oct. 2020.

[9] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral-spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.

[10] L. Fu *et al.*, "Learning robust discriminant subspace based on joint $L_{2,p}$- and $L_{2,s}$-norm distance metrics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 130–144, Jan. 2022.

[11] Q. Ye, P. Huang, Z. Zhang, Y. Zheng, L. Fu, and W. Yang, "Multiview learning with robust double-sided twin SVM," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2021.3088519.

[12] 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.

[13] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[14] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.

[15] E. Irwansyah, Y. Heryadi, and A. A. S. Gunawan, "Semantic image segmentation for building detection in urban area with aerial photograph image using U-Net models," in *Proc. IEEE Asia-Pacific Conf. Geosci., Electron. Remote Sens. Technol.*, 2020, pp. 48–51.

[16] Z. Zhang, C. Zhang, and W. Li, "Semantic segmentation of urban buildings from VHR remotely sensed imagery using attention-based CNN," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1833–1836.

[17] C. Tian, C. Li, and J. Shi, "Dense fusion classmate network for land cover classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 262–2624.

[18] X. Tan, Z. Xiao, Q. Wan, and W. Shao, "Scale sensitive neural network for road segmentation in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 533–537, Mar. 2021.

[19] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.

[20] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2612–2626, Aug. 2019.

[21] X. Zhang, Z. Xiao, D. Li, M. Fan, and L. Zhao, "Semantic segmentation of remote sensing images using multiscale decoding network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1492–1496, Sep. 2019.

[22] J. Liu, X. Xiong, J. Li, C. Wu, and R. Song, "Dilated residual network based on dual expectation maximization attention for semantic segmentation of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1825–1828.

[23] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[24] D. Lei, T. Hao, and B. Lorenzo, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.

[25] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2022.3164083.

[26] L. Sun, F. Wu, C. He, T. Zhan, W. Liu, and D. Zhang, "Weighted collaborative sparse and $L_{1/2}$ low-rank regularizations with superpixel segmentation for hyperspectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art no. 5500405.

[27] C. He, L. Sun, W. Huang, J. Zhang, Y. Zheng, and B. Jeon, "TSLRLN: Tensor subspace low-rank learning with non-local prior for hyperspectral image mixed denoising," *Signal Process.*, vol. 184, 2021, Art. no. 108060.

[28] Y. Zheng, B. Jeon, L. Sun, J. Zhang, and H. Zhang, "Student's T-hidden Markov model for unsupervised learning using localized feature selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2586–2598, Oct. 2018.

[29] X. Wang, C. Shen, H. Li, and S. Xu, "Human detection aided by deeply learned semantic masks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2663–2673, Aug. 2020.

[30] 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.

[31] 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.

[32] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1857–1866.
[33] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2015. [Online]. Available: http://arxiv.org/abs/1412.7062
[34] L.-C. Chen, G. Papandreou, and I. Kokkinos, "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.
[35] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.05587
[36] 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.
[37] 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.
[38] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018.
[39] D. Lei, Z. Jing, and B. Lorenzo, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5367–5376, Aug. 2020.
[40] B. Cui, W. Jing, L. Huang, Z. Li, and Y. Lu, "SANet: A sea–land segmentation network via adaptive multiscale feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 116–126, 2021.
[41] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[42] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
[43] 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.
[44] J. Park, S. Woo, J.-Y. Lee, and I.-S. Kweon, "BAM: Bottleneck attention module," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1807.06514
[45] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.
[46] J. Huang, X. Zhang, Y. Sun, and Q. Xin, "Attention-guided label refinement network for semantic segmentation of very high resolution aerial orthoimages," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4490–4503, 2021.
[47] Z. Zheng, X. Zhang, P. Xiao, and Z. Li, "Integrating gate and attention modules for high-resolution image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4530–4546, 2021.
[48] L. Mou, Y. Hua, and X. X. Zhu, "Spatial relational reasoning in networks for improving semantic segmentation of aerial images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5232–5235.

Le Sun (Member, IEEE) was born in Jiangsu, China, in 1987. He received the B.S. degree in mathematics from the School of Science, Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2009, and the Ph.D. degree in computer science from the School of Computer Science and Engineering, NJUST, in 2014.

From 2015 to 2018, he was a Postdoctoral Researcher with the School of Electronic and Electrical Engineering, Sungkyunkwan University, Seoul, South Korea, where he conducted research in the field of multi-image fusion based on sparse dictionary learning and compressive sensing. Since 2020, he has been an Associate Professor with the School of Computer and Science, Nanjing University of Information Science and Technology, Nanjing. His research interests include hyperspectral image processing (including unmixing, classification, and restoration), sparse representation, compressive sensing, and deep learning.

Shiwei Cheng received the B.S. degree in software engineering from the School of Software College, Zhongyuan University of Technology, Zhengzhou, China, in 2019. He is currently working toward the M.S. degree in electronic information with the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include semantic segmentation of remote sensing images.

Yuhui Zheng (Member, IEEE) was born in Shanxi, China, in 1982. He received the B.S. degree in chemistry and the Ph.D. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2004 and 2009, respectively.

From 2014 to 2015, he was a Visiting Scholar with the Digital Media Laboratory, School of Electronic and Electrical Engineering, Sungkyunkwan University, Seoul, South Korea. He is currently a Full Professor with the School of Computer and Science, Nanjing University of Information Science and Technology, Nanjing. His research interests include image processing, pattern recognition, and remote sensing information systems.

Zebin Wu (Senior Member, IEEE) was born in Zhejiang, China, in 1981. He received the B.S. and Ph.D. degrees in computer science from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2003 and 2008, respectively.

He is currently a Full Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include virtual reality and system simulation, remote sensing information processing, and distributed computing.

Jianwei Zhang received the B.S. degree from Wuhan University, Wuhan, China, in 1998, and the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2006, both in computer science.

He is currently a Professor with the College of Mathematics and Physics, Nanjing University of Information Science and Technology, Nanjing. His research interests include pattern recognition, artificial intelligence, and remote sensing information processing.