SDCAFNet: A Deep Convolutional Neural Network for Land-Cover Semantic Segmentation With the Fusion of PolSAR and Optical Images

Boce Chu, Jinyong Chen, Jie Chen, *Senior Member, IEEE*, Xinyu Pei, Wei Yang, *Member, IEEE*,
Feng Gao, and Shicheng Wang

***Abstract*—Due to the different imaging mechanisms between optical and polarimetric synthetic aperture radar (PolSAR) images, determining how to effectively use such complementary information has become an interesting and challenging problem. Convolutional neural networks (CNNs) and other deep neural networks have achieved good experimental results in remote sensing land-cover semantic segmentation. However, the CNN convolution structure can extract only the features within the receptive field in the spatial dimension without focusing on the relationship between multiple channels; therefore, it is impossible to realize fusion and complementarity between multiple channels. In this article, we propose a novel spatial dense channel attention fusion network (SDCAFNet), which takes optical and PolSAR images as different inputs and completes feature fusion and semantic segmentation within a neural network. First, SDCAFNet uses a two-stream siamese CNN network to realize the preliminary feature coding of optical and PolSAR images. Then, a spatial dense channel attention module (SDCAM) is proposed. The channel activation values obtained at different positions are combined in the spatial dense matrix, which can describe the attention in the feature fusion process. Finally, we introduce the fused features into the symmetric skip-connection decoder composed of multiple symmetric decoder blocks to realize end-to-end land-cover semantic segmentation. Experimental results show that SDCAFNet can effectively learn the correlation between optical and PolSAR channels and has a better segmentation accuracy than other methods.**

***Index Terms*—Channel attention, feature fusion, land-cover semantic segmentation, optical image, polarimetric synthetic aperture radar (PoLSAR).**

Manuscript received 21 December 2021; revised 18 June 2022; accepted 6 October 2022. Date of publication 11 October 2022; date of current version 21 October 2022. This work was supported by the S&T Program of Hebei under Grant 21340302D and Foundation of the CETC Key Laboratory of Aerospace Information Applications under Grant SCX20629T007. *(Corresponding author: Boce Chu.)*

Boce Chu is with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China, and also with the Key Laboratory of Aerospace Information Applications of CETC, Shijiazhuang 050081, China (e-mail: bocc012628077@163.com).

Jinyong Chen, Xinyu Pei, Feng Gao, and Shicheng Wang are with the Key Laboratory of Aerospace Information Applications of CETC, Shijiazhuang 050081, China (e-mail: by1902077@buaa.edu.cn; 938566810@qq.com; 19903113711@126.com; wang.shicheng@163.com).

Jie Chen and Wei Yang are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: chenjie@buaa.edu.cn; yangweigigi@ee.buaa.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3213601

I. Introduction

LAND-COVER semantic segmentation has long been a fundamental but challenging research topic in geoscience and remote sensing (RS). At present, many RS semantic segmentation tasks use data from a single sensor. Many studies use optical RS images for semantic segmentation, but the corresponding effect is poor due to unitary spectral information and cloud cover [1], [2]. Other studies use polarimetric synthetic aperture radar (PolSAR) images for semantic segmentation [3], [4], but the corresponding semantic segmentation effect is also poor because of speckle noise, which affects the performance of pixel-based polarimetric decomposition features. With the development of artificial intelligence, many advanced deep or machine learning methods, such as GCNs [5] and transformer mechanisms [6], have been applied in single-modal RS classification, which can achieve more robust optimum and better detailed spectral representations.

Additionally, studies [7] have shown that the ability to identify materials on the surface of the Earth remains limited due to the lack of rich and diverse information, particularly in challenging scenes where certain categories are similar and cannot be accurately classified by only single modalities [8], [9], [10]. Different imaging technologies in RS are capable of capturing a variety of properties from the Earth's surface [11], such as spectral radiance and reflectance, height information, texture structure, and spatial characteristics. The joint exploitation of multiple modalities enables us to characterize the scene at a more detailed and precise level unachievable using single modality data [12].

With the rapid development of RS technology, such as satellite constellations and satellites with multiple sensors, it is possible to obtain images in the same area from a variety of sensors. It is possible to improve the segmentation accuracy by fusing multisource image data. There are great differences in the geometric and radiation characteristics between PolSAR images and optical images. The accuracy of land-cover semantic segmentation can be improved by fusing high spatial resolution optical and PolSAR images [13], [15]. Therefore, designing a suitable deep or machine learning model to complete the advantages complementary to optical and PolSAR images is an urgent problem to solve [16], [17].

There are three issues to be solved in land-cover semantic segmentation with the fusion of PolSAR and optical, namely, the selection of semantic segmentation methods, the feature extraction methods of optical and PolSAR and the fusion methods of multimodal features.

In terms of semantic segmentation algorithm selection, there are two mainstream methods: the classification method based on image patches [18] and the end-to-end segmentation network [19]. The first method requires the division of the RS image into patches, and each patch needs to be classified separately. There are already some good methods to improve the classification accuracy of each patch by designing a new neural network structure [7], [20], which effectively improves the classification results of multimodal fusion.

The representative of the second method is the full convolution neural network, which was proposed by Long et al. [21]. The segmentation results with the same size as the input image can be directly obtained through a deconvolution operation. The full convolutional neural network (CNN) and its deformation networks, such as UNet [22] and DeepLab [23], have gradually become mainstream in semantic segmentation tasks because of their ability to input images of any size and achieve high-precision classification results. At present, most of the existing fully CNNs only support single-mode image classification. In this article, we creatively propose an end-to-end semantic segmentation network spatial dense channel attention fusion network (SDCAFNet), which can support both inputs of optical and PolSAR images and complete semantic segmentation efficiently.

In terms of the feature extraction methods selection, the existing feature extraction method for image fusion mostly adopts the method based on multimodal deep learning models. There are two mainstream multimodal model methods. One method is that each model uses the same network to extract features [24], and the other method is that their feature networks are designed for different modalities [25]. The first method cannot extract the optimal features according to the imaging characteristics of multisource images because the feature extraction networks are the same, which reduces the feature diversity and description power. The second method can design the most appropriate network structure to extract features according to the image characteristics of different modes; this structure is better than that of the first method. In this article, we used a two-stream siamese feature encoder in SDCAFNet, which is used as the feature extraction network for optical and PolSAR images without weight sharing. It is worth mentioning that to maximize the feature extraction ability of each stream encoder, we use a large number of optical images and PolSAR images to pretrain each stream encoder separately.

In terms of the multimodal feature fusion method, the attention mechanism, especially the channel attention mechanism, has been applied to the field of computer vision in recent years [26]. In the image classification task, the channel attention mechanism aims to obtain the contribution relationship between multiple channels and enhance the final classification efficiency by activating or suppressing different channels. A squeeze-and-excitation network (SENet) [27], which is the most representative method, converts the global spatial information extracted through the global average pool into a multilayer perceptron and finally generates an attention map to describe the relationships between channels. However, different from the image-wise classification task, the land-cover classification task belongs to the domain of pixel-wise segmentation, and the relationships between channels in each local range of the image are different. The global attention obtained by SENet cannot describe the local differentiated attention. To solve the abovementioned problems, we propose the spatial dense channel attention module (SDCAM) in SDCAFNet, which can record the local correlation between channels in the spatial dense matrix and introduce a refined local attention mechanism in disguise.

In conclusion, this article proposes a new deep learning model, SDCAFNet, to realize end-to-end land-cover semantic segmentation through the fusion of optical and PolSAR images. SDCAFNet consists of three components: a two-stream siamese feature encoder, feature fusion with SDCAM and symmetric skip-connection decoder. First, we propose a two-stream CNN feature encoder and pretrained encoders of each stream with a large number of labeled PolSAR and optical images to improve the feature extraction ability in multimodal RS images to the greatest extent. Then, the SDCAM is proposed. The nonlinear relationships of the optical and PolSAR channels at each local position are extracted through the spatial dense matrix, and the local attention map is obtained. The attention map is combined with the multimodal features of the encoder output to realize the nonlinear optimization of the features. After that, we realize the feature fusion of optimized features through convolution. Finally, we propose a symmetric skip-connection decoder to obtain semantic segmentation in an end-to-end manner. Additionally, the decoder can skip-connect the fused features with the shallow features of the optical and PolSAR images. By making full use of the original spatial information in the optical and PolSAR images, the upsampling precision is improved to the greatest extent.

The main contributions of this article are as follows.

First, we propose a new network structure called SDCAFNet, which can better complete the end-to-end land-cover segmentation task with the fusion of optical and PolSAR features. Unlike the patch-wise multimodal fusion networks, SDCAFNet can output the classification results of the original image size without patch-by-patch calculations. We design the network structure using some full convolution networks for reference.

Second, we propose a feature selection and fusion module called SDCAM, which can obtain the local relationships between channels by applying a reasonable compression ratio to the spatial dense matrix structure. Additionally, through channel selection at different locations, the dense fused features of optical and PolSAR images are obtained by SDCAM, which can effectively solve the feature space inconsistency problem of multimodal images.

Third, we also propose a symmetric decoder block that is suitable for end-to-end semantic segmentation tasks with multisource fusion. This block can simultaneously introduce the high-resolution contour features of optical and PolSAR in the decoding process, which can combine the rich spatial location information in the low-level features with the rich semantic information in the high-level features.

II. Related Work

A. Land-Cover Semantic Segmentation Based on Optical and PolSAR Fusion

There are several methods for improving the performance of land-cover classification by fusing SAR and optical images [28], [29]. However, there are great differences in the geometric and radiometric characteristics between optical and SAR images. In particular, PolSAR images contain more information than single-band SAR images [30], [31], [32]. At present, there is no better method for realizing the effective feature fusion of optical and PolSAR images. Many studies [33], [34] attempt to solve the land-cover classification task through a fusion strategy after differential feature extraction for optical and SAR data. The most important part of this process is the feature extraction of multimodal data and classifier design. The Siamese CNN [25], [35] uses the powerful CNN function to derive the high-level features of multimodal datasets and then concatenates these features for classification. Specifically, concatenation, which is an effective and simple method, is the main strategy for fusing PolSAR and optical data thus far. In essence, SDCAFNet also draws lessons from the strategy of feature concatenation. However, the difference is that we introduce the local attention mechanism into the concatenation process so that the fused features are more suitable for semantic segmentation tasks.

B. Attention and Gating Mechanisms

Recently, attention mechanisms have received great attention. Attention is a method for concentrating computing resources on the most valuable part of the network for classification tasks. The effectiveness of attention has been proven in many kinds of applications [36], [37], [38], [39]. Chen et al. [40] proposed a multiscale feature attention mechanism for semantic segmentation tasks. Wang et al. [41] introduced a powerful bottom-up, top-down feedforward attention mechanism based on hourglass modules that are inserted into deep residual networks. Hu et al. [27] proposed a channel attention graph to identify the global relationship between channels through convolution. In contrast, our proposed SDCAM can be considered a denser local attention mechanism.

III. Methodology

We propose a network structure SDCAFNet, which can improve the land-cover semantic segmentation result through the selection and fusion of the feature channels of optical and PolSAR images. The architecture of SDCAFNet is shown in Fig. 1, which consists of three components: a two-stream siamese feature encoder, SDCAM, feature fusion and symmetric skip-connection decoder. Different from previous fusion networks, such as the M3 [42] and binary complex neural network [43], SDCAFNet is an end-to-end fusion network for semantic segmentation, and the input of the network is the original image rather than the segmented patches. First, the original optical and PolSAR images obtain the corresponding initial features $X \in R^{C\times H\times W}$ and $Y \in R^{C\times H\times W}$ through the two-stream siamese feature encoder. Then, X and Y are used as the inputs of the SDCAM to obtain the optimized features $\overline{U} \in R^{2C\times H\times W}$ through local attention channel selection. Finally, we fuse the optimized features through a convolution operation and directly obtain the classification result graph with the same input size through a symmetric skip-connection decoder.

In this section, we describe the implementation process of the proposed SDCAFNet in detail. In Section III-A, we introduce the network structure of the two-stream siamese feature encoder and the pretraining method in detail. Then, in Section III-B, we introduce the structure and principle of the SDCAM in detail. Finally, we introduce the network structure of the symmetric skip-connection decoder and the training process of SDCAFNet in Section III-C.

A. Two-Stream Siamese Convolution Feature Encoder

Due to the different imaging mechanisms for optical and PolSAR images, the visual effects of the two types of images are quite different, and the internal features are not in the same feature space. To achieve effective fusion, different modal image features need to be transformed into the same feature space through differentiated feature encoding. A neural network can obtain the abstract expression of input data in high dimensional space. Different network parameters represent different feature mappings. We use two full convolution networks that do not share parameters to extract independent features from optical and PolSAR images and map them to the same high-dimensional feature space to provide a spatial basis for the subsequent fusion process.

The network structure of the two-stream siamese convolution feature encoder is shown in Fig. 2. PolSAR and optical images are encoded through a series of convolution layers, in which we use the smaller scale 3×3 instead of using a larger convolution filter. Considering that a small convolution filter increases the nonlinearity inside the network, the network has a relatively strong discrimination ability [25]. The 3×3 convolution filter is the smallest kernel for capturing different directional modes, such as center, up and down, left and right, so we choose a 3×3 convolution filter to complete the encoding process. Max pooling is performed over 2×2 pixel windows with a stride of 2. All layers in the network are equipped with a nonlinear linear unit (ReLU) as an activation function. The last layer of the encoder is connected by a series of deconvolution networks in the pretraining process and will be connected by the SDCAM in the following training process.

To improve the training efficiency of SDCAFNet, we propose a reasonable pretraining method for the encoder. By maximizing the quantity of pretrained image data, the encoder can maximize the differential feature extraction ability in optical and PolSAR images. By referring to the UNet [22], we propose two decoders for each of the two-stream encoders, which form two independent UNets, PolSAR-UNet and optical-UNet. Then, the pretraining task of the encoder in SDCAFNet is converted into the training tasks of PolSAR-UNet and optical-UNet. It is

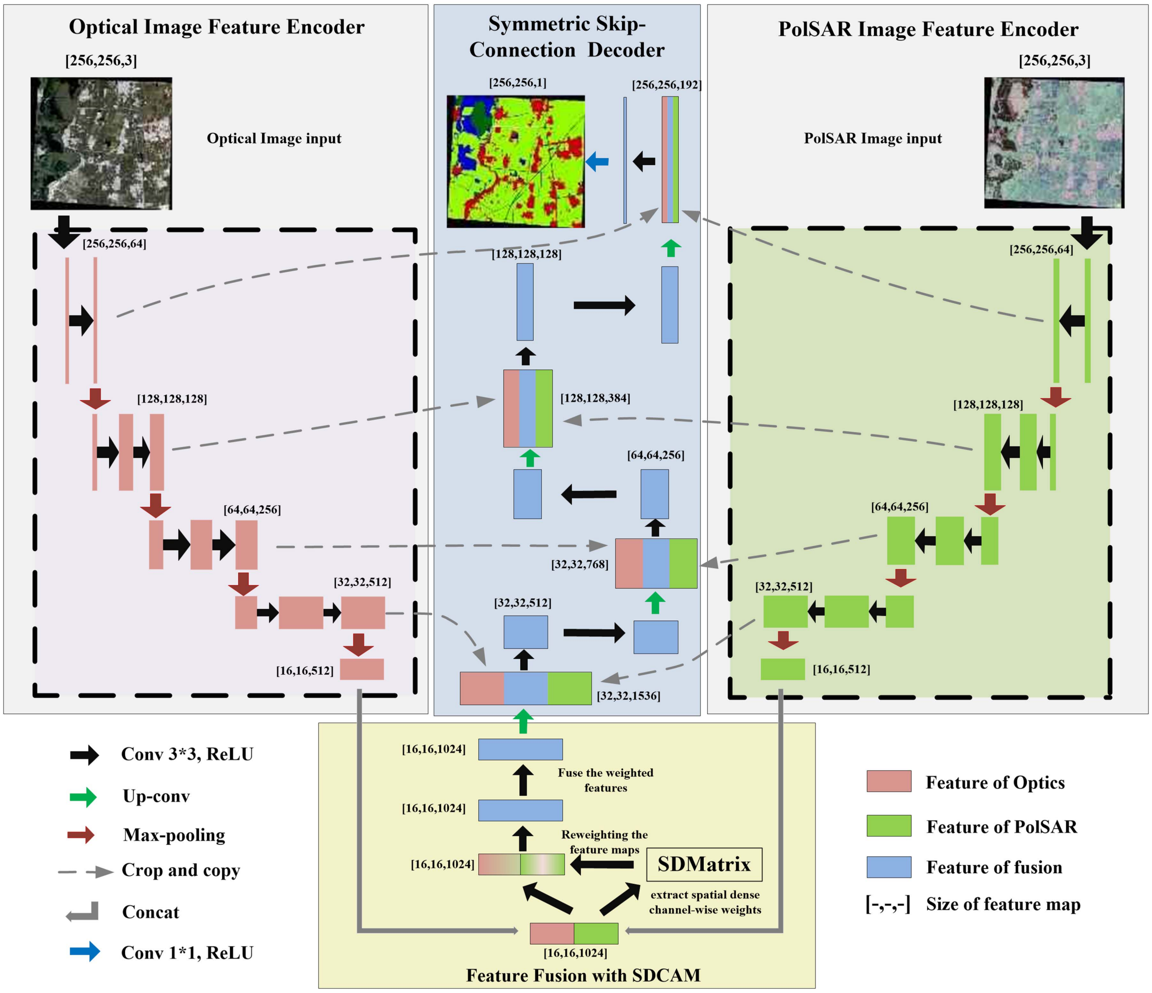


Fig. 1. Flowchart of SDCAFNet.

worth mentioning that the training process of PolSAR-UNet and optical-UNet is independent, so the data available for training are not limited to optical and PolSAR images with overlapping areas. This will greatly reduce the constraints on the available data for training.

Most of the existing image classification studies use ImageNet as the pretraining dataset [44]. However, ImageNet is mostly normal photos, which are quite different with respect to the observation angle and imaging performance of RS images, and the network parameters pretrained by ImageNet are not optimal. In this article, we first make full use of ImageNet to preliminarily train the optical-UNet and PolSAR-UNet network parameters so that the network can obtain the basic image semantic extraction ability. Then, a large number of optical RS images are used for further migration training based on the preliminarily trained optical-UNet so that the network can obtain a feature extraction ability that is suitable for the characteristics of optical images. Through the abovementioned method, the trained optical-UNet encoder can be used as the pretraining result of the optical-stream encoder in SDCAFNet. Similarly, the PolSAR-stream encoder in SDCAFNet can also be obtained in the above way. The advantage of this is that it can make maximum use of the current massive amounts of open-source photos and RS image data, ensure that the network parameters of the encoder are more suitable for the characteristics of RS images, and provide a better network parameter basis for further training optical and PolSAR fusion parameters for SDCAFNet.

B. Spatial Dense Channel Attention Module

As a new attention mechanism module, our SDCAM can obtain the dense relationship between channels at different local positions based on a spatial dense matrix. SDCAM is a computing block based on the feature output of the two-stream siamese convolution feature encoder. In the notation that follows, we take $X \in R^{(H\times W\times C)}$ to be the PolSAR-stream feature and $Y \in R^{(H\times W\times C)}$ to be the optical-stream feature. We can then copy and concatenate X and Y by channels as $U \in R^{(H\times W\times 2C)}$, which are directly applied as the input to the subsequent feature compression and channel relation extraction. SDCAM consists of two steps: feature compression by local information embedding and channel relation extraction of local spatial information. A diagram illustrating the structure of SDCAM is shown in Fig. 3.

1) Feature Compression by Local Information Embedding: Generally, the relationship between the channels at each location

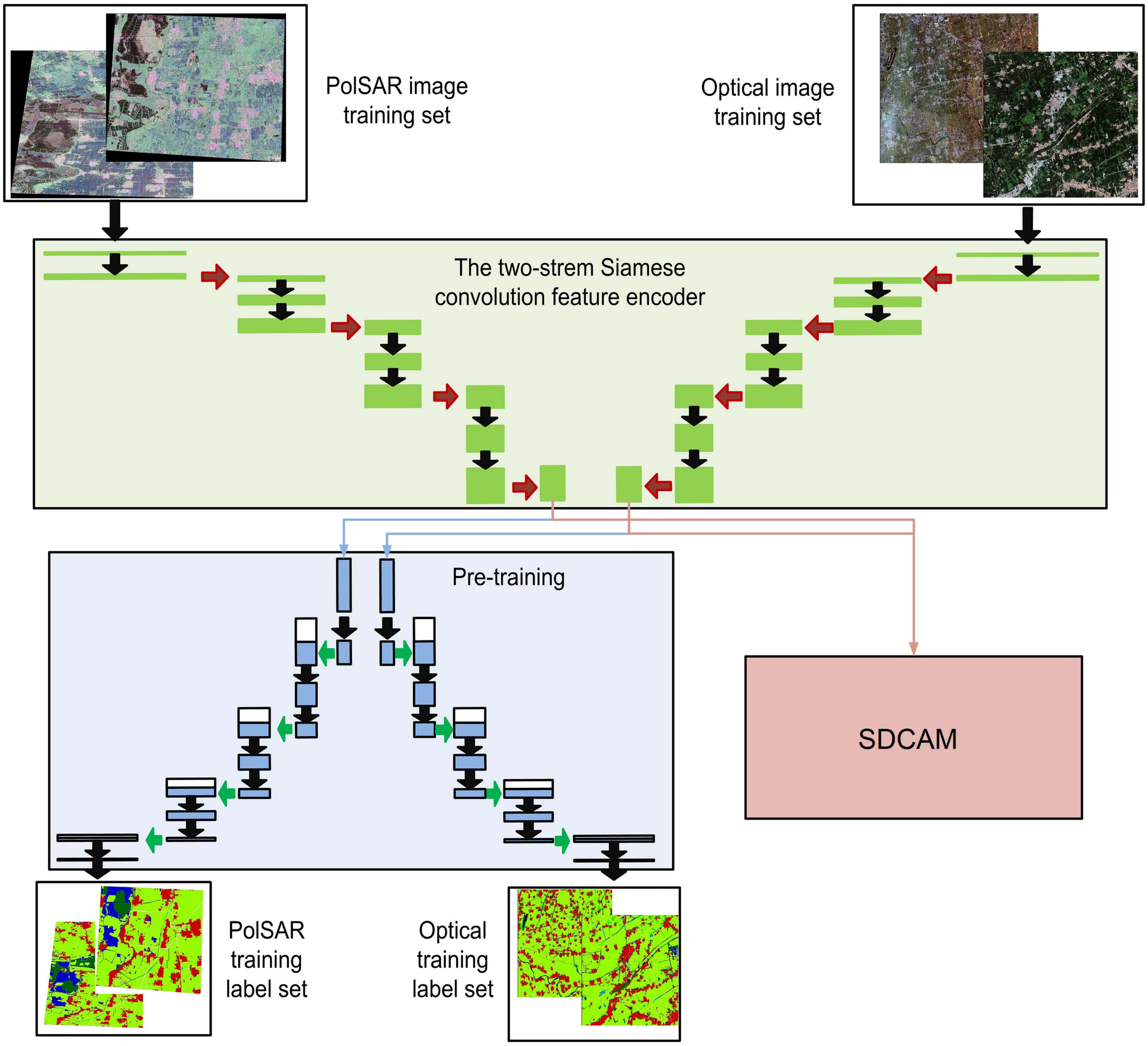

PolSAR image training set
Optical image training set
The two-strem Siamese convolution feature encoder
Pre-training
SDCAM
PolSAR training label set
Optical training label set

Fig. 2. Network structure and pretraining process of the two-stream convolution feature encoder.

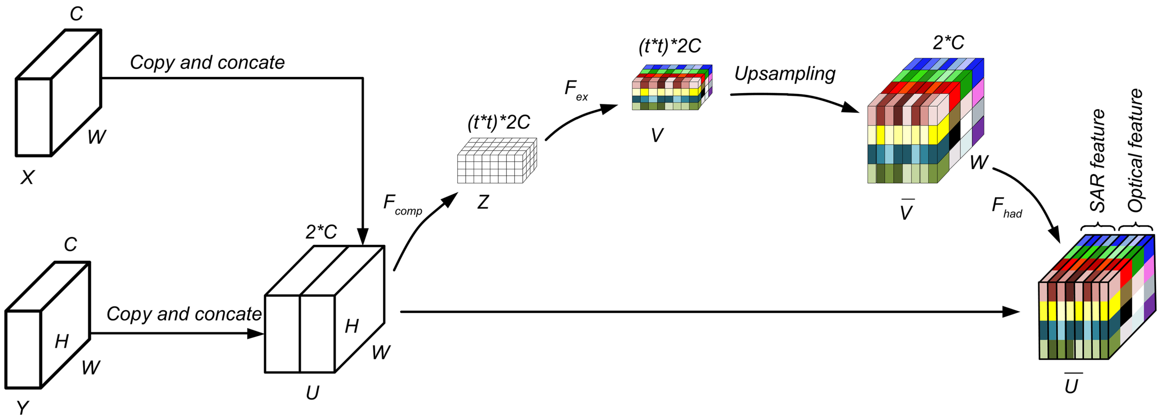


Fig. 3. Details of the SDCAM.

is not only related to the channels at that location but is also affected by the channel information in the surrounding adjacent area. Therefore, it is necessary to make effective use of contextual information in calculating the relationship between channels within a local location. To effectively extract the dense relationship between the channels, we intend to compress the features locally by embedding local information and use the statistical feature to represent the features of the area. Additionally, the land-cover distribution of RS images is continuous, and most of the land cover in the local spatial area is approximate.

In other words, in the local area, the statistical features and the relationships between channels are similar. Therefore, on the premise of reasonable selection of the compression ratio, the compression of local features will not cause the problem in which the calculation results of the statistical features and the local relationships between channels are not representative. In summary, using feature compression by local information embedding can not only effectively introduce contextual information to make the calculation between channels more reasonable but also effectively reduce the parameter dimension and improve the calculation efficiency without introducing more errors.

To compress the local spatial information of the image, we design an operation F_{comp}, which takes the statistical average pixel as the descriptor of the region to describe the local spatial information in the channel. A reasonable feature compression ratio *t* in F_{comp} should be set according to the size of the input image. When the input image is larger, the compression ratio *t* can be appropriately increased. By constantly adjusting the value of *t*, we can realize the local description with the most reasonable granularity. The local spatial information descriptor at the (*m*, *n*) position in *Z* can be obtained in the following way:

$$\begin{aligned} Z_{(m,n)} &= F_{\text{comp}}(U,m,n) \\ &= \frac{\sum_{i=1+(m-1)\frac{H}{t}}^{\frac{m*H}{t}} \sum_{j=1+(n-1)\frac{H}{t}}^{\frac{n*W}{t}} U(i,j)}{\frac{H}{t} * \frac{W}{t}}. \end{aligned} \tag{1}$$

In Fig. 3, the compression ratio $t = 4$, which can compress $U \in R^{(H\times W\times 2C)}$ to $Z \in R^{(4\times 4\times 2C)}$.

2) Local Spatial Channel Relationship Extraction: The ultimate goal of SDCAM is to capture the relationships between channels and apply them to feature channel selection. In the first step, we compress the information to the feature descriptor of each local region. In this step, we need to make full use of the feature descriptor for further training to learn the complex nonlinear relationship between channels at different local positions. We design a gating mechanism to learn the nonlinear relationship between channels by imitating long short-term memory [45], gated recurrent units [46], and other recurrent neural network-related networks [47]. The formula is defined as follows:

$$V = F_{\text{ex}}(Z, W_{\text{fc}_1}, W_{\text{fc}_2}) = W_{\text{fc}_2}(\text{sigmoid}(W_{\text{fc}_1} Z)) \tag{2}$$

where $W_{\text{fc}_1} \in R^{(\frac{2C}{r}\times 2C)}$ and $W_{\text{fc2}} \in R^{(2C\times\frac{2C}{r})}$. We parameterize the gating mechanism by forming two fully connected layers (fc_1 and fc_2) and a nonlinear activation function (sigmoid), which can not only reduce the network complexity but also flexibly learn the nonlinear relationship between channels. Through gating mechanism training, we can obtain the spatial dense matrix. Then, the spatial dense matrix is upsampled, and its size is set equal to *U* before feature compression. Finally, the spatial dense matrix is taken as the Hadamard with *U* to complete the feature channel selection. The formula is as follows:

$$\bar{U} = F_{had}(U, V) = \bar{V} \times U = Upsampling(V) \times U. \tag{3}$$

The overall processing of the proposed SDCAM is shown in Algorithm 1.

Algorithm 1: The Overall Processing of SDCAM.

```
Input: The feature maps X and Y extracted by the
       two-stream convolution encoder;
Output: The feature maps weighted by the spatial dense
        matrix;
  1. Concatenate the feature maps X and Y as a whole
     feature map U;
  2. Compress feature map U to embedded feature map Z
     based on Eqs. (1);
  3. Extract the channel relationship descriptor V, also
     called the spatial dense matrix, by the gating
     mechanism based on Eq. (2);
  4. Upsample the V into $\overline{V}$, which is the same size as U;
  5. Weight U into $\overline{U}$ by $\overline{V}$ based on Eq. (3), which can
     introduce the spatial channel-wise attention
     operation into the feature maps.
```

It can be determined from the above steps that SDCAM can obtain spatial local attention through training and optimize the features of optical and PolSAR images. Next, a feature fusion block, which is represented by (4), is used to fuse the weighted features of optical and PolSAR images to obtain the fused feature *Q*

$$Q = W_2\delta(W_1 * \bar{U}) \tag{4}$$

where W_1 symbolizes the first convolution operation, W_2 symbolizes the second convolution operation, and δ illustrates the ReLU function.

C. Symmetric Skip-Connection Decoder

To realize the end-to-end land-cover classification network after obtaining the fused features, we need a decoder to convert the features into land-cover classes. Considering that the input consists of optical and PolSAR images and the detailed information in both image modalities is useful for recovering spatiotemporal information and improving resolution, we propose the symmetric skip-connection decoder, which can connect the fused features with the low-level features of optical and PolSAR images. The symmetric skip-connection decoder is composed of multiple decoder blocks. The detailed structure of the decoder block is shown in Fig. 4.

The first layer of the decoder block is a concatenated feature map named C_k

$$C_{\text{k}} = [X_k, Y_k, Z_k] \tag{5}$$

where *k* refers to the serial number of the decoder block, X_k is the feature from the PolSAR-stream encoder, and Y_k is the feature from the optical-stream encoder. Z_k is the output of the previous decoder block. It is worth mentioning that when $k = 1$, Z_k is equal to the fused feature *Q*.

In addition to the concatenated feature map C_k, the decoder block also contains two convolutional layers and an up-conv layer. In the last decoder block, we remove the up-conv layer and directly take the output of two convolutional layers as the final classification result.

TABLE I
SUMMARIZED STATISTICS OF THE BAIYANGDIAN DATASET

Class	Total number of pixels	Number of the pixels in the training dataset	Number of the pixels in the validation dataset	Number of the pixel in the test dataset
W	7 868 173	5 193 612	927 412	1 747 149
Bu	4 926 134	3 917 533	282 913	725 688
T	861 023	485 718	197 452	177 853
F	11 335 333	8 912 175	1 084 672	1 338 486
G	207 284	103 742	41 837	61 705
Ba	5 139 272	3 048 372	1 037 462	1 053 438
Un	30 481	28 171	291	2019
Total	30 367 700	21 689 323	3 572 039	5 106 338

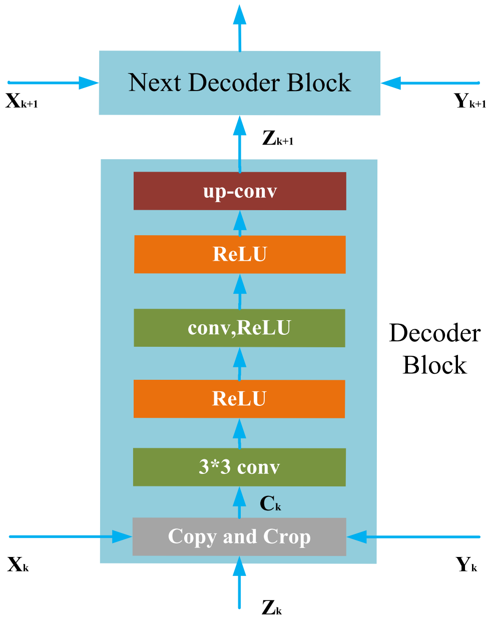


Fig. 4. Details of the symmetric skip-connection decoder block.

In summary, the detailed structure of SDCAFNet is provided in Fig. 5.

IV. Experiment Evaluation

In this section, we conduct several experiments to evaluate the land-cover semantic segmentation performance of SDCAFNet. First, the datasets and experimental settings used in the experiment are introduced in Section IV-A. The comparison methods are introduced in Section IV-B. The configuration of important hyperparameters, such as input size and compression ratio, is introduced in Section IV-C. In Section IV-D, we compare the experimental results for two datasets between SDCAFNet and the comparison methods. Section IV-E shows the validation curves of the two datasets. Section IV-F shows the computational time. Finally, Section V concludes this article.

A. Dataset and Experimental Settings

The Xiong'an New Area is a special economic zone that has been established in China in recent years. Its land-cover distribution and change are of greater concern to society and have high RS Earth observation value. Additionally, the land-cover classes in Xiong'an New Area are complex, including cities and wetland rivers, which can effectively verify the universality of the algorithms. Therefore, we conduct several experiments to evaluate the land-cover semantic segmentation performance of SDCAFNet on two datasets that can cover different geomorphic features of the Xiong'an New Area.

The first coregistered optical and PolSAR image datasets are located in the Baiyangdian Wetland Nature Reserve, which has typical wetland coverage; the dataset size is 5812 × 5225, with a spatial resolution of 3.2 m. Table I lists the pixel numbers of training, validation, and testing samples for each class in the Baiyangdian dataset.

The second coregistered optical and PolSAR image dataset are located in Rong County, Hebei Province, where typical urban land cover exists; the dataset size is 3449 × 6410, with a spatial resolution of 3.2 m. We call this dataset "Rong." Table II lists the pixel numbers of the training, validation, and testing samples for each class in the Rong dataset.

The optical images in the two datasets are obtained from GaoFen-2 in multispectral bands, as shown in Fig. 6(a). PolSAR images are GaoFen-3 level 1A products in the *C*-band and quad-polarization (HH + HV + VH + VV) observation modes, as shown in Fig. 6(b). Because the resolution of GaoFen-2 is different from that of GaoFen-3, to maintain a consistent resolution, we resized the GaoFen-3 image, which has a higher resolution, to the GaoFen-2 size. We use a deep generative matching network [48] to register optical and PolSAR images. Then, a manual fine adjustment is used to obtain the final coregistered optical and PolSAR image dataset used for fusion. In addition, to calculate and compare the indexes of different methods, we labeled the ground truth of the Baiyangdian and Rong datasets according to the visual interpretation of the optical images. All the pixels are assigned to seven classes: building (Bu), farmland (Fa), tree (T), grass (G), bare land (Ba), water (W), and unknown (Un). The ground truth is shown in Fig. 6(c).

The settings of our experiment are as follows. Learning rates are tested between $(0, 10^{-8})$. Through multiple experiments on two datasets, 10^{-6} is selected as the final learning rate, which can achieve rapid convergence and avoid large amplitude oscillations in the gradient descent. The loss function is the weighted categorical cross-entropy [48]

$$\text{Funtion}_{\text{loss}} = -\sum_{i=1}^{k} W_i[y_i \log \overline{y_i} + (1-y_i)\log(1-\overline{y_i})] \quad (6)$$

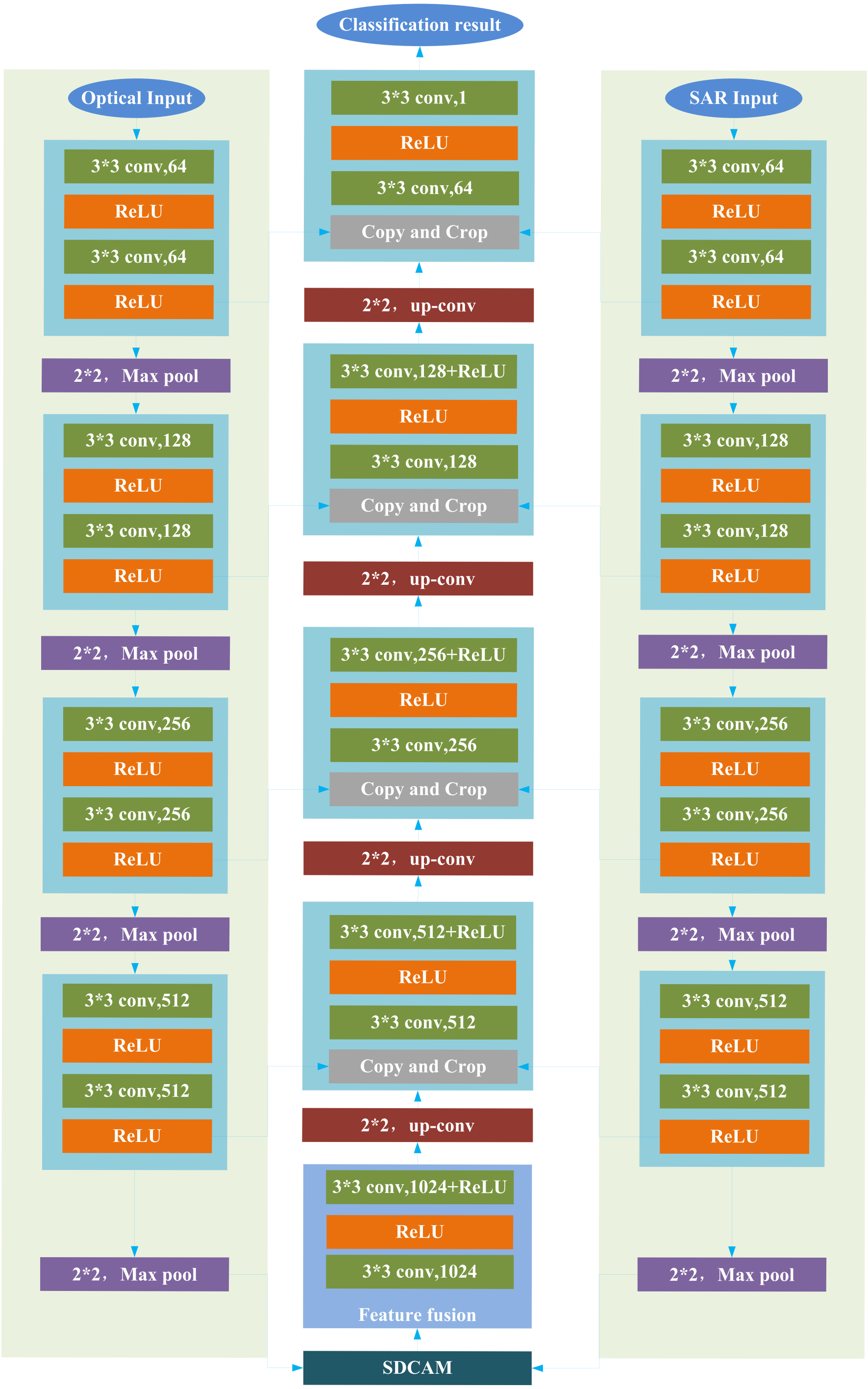


Fig. 5. Structural details of the symmetric skip-connection decoder.

where y_i is the ground truth with one-hot coding for class i, $\overline{y_i}$ is the softmax function output for class i, k is the total number of output classes, and W_i is the balanced weights of class i.

We set the epoch to infinity, and the training termination condition is met only when the loss function is less than 10^{-3} and the difference between two consecutive losses is smaller than 10^{-4}. Considering the limited memory, the batch size is set to 10 in the experiments. All experiments are executed in TensorFlow and Python 3.5 on the Windows platform with an NVIDIA Quadro P5000 GPU (16 GB), and we also use the OpenCV toolbox to carry out the image preprocessing work.

B. *Comparison Methods*

UNet+Optical: Recently, UNet has achieved good results in the semantic segmentation of single-modal RS images.

TABLE II
SUMMARIZED STATISTICS OF THE RONG DATASET

Class	Total number of pixels	Number of the pixels in the training dataset	Number of the pixels in the validation dataset	Number of the pixel in the test dataset
W	4 578 010	3 055 931	602 149	919 930
Bu	4 378 645	3 733 551	209 400	435 694
T	1 363 506	1 013 912	159 591	190 003
F	11 087 935	6 011 908	1 673 095	3 402 932
G	667 245	525 307	71 544	70 394
Ba	24 834	16 149	7485	1200
Un	7915	7296	306	313
Total	22 108 090	14 364 054	2 723 570	5 020 466

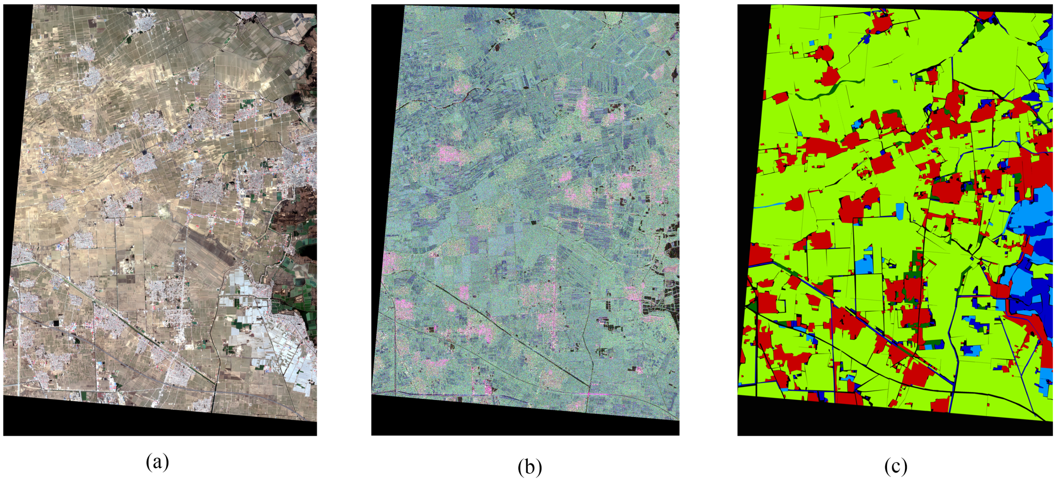


Fig. 6. Examples of the optical, PolSAR (Pauli-RGB) and ground truth images. (a) Optical image. (b) PolSAR (Pauli-RGB) image. (c) Ground truth.

To compare the performances of single-modal methods and multimodal fusion methods, we chose UNet as the comparison method of optical image semantic segmentation. We pretrained on ImageNet and finetuned the optical images of the Baiyangdian and Rong datasets. We refer to this method as "UNet-Op."

UNet+PolSAR: This method is similar to the construction of "UNet-Op." The only difference is that the input is changed from optical to PolSAR images. We refer to this method as "UNet-PS."

UNet+Optical+PolSAR: We concatenated the optical images and PolSAR images into eight channels (R+G+B+NIR+HH+HV+VH+VV) as the UNet input. We refer to this method as "UOP."

UNet+Optical+ PolSAR+SE: The squeeze-and-excitation (SE) block [22] is embedded into the UOP method, which can introduce a global attention mechanism into the feature encoder. We refer to this method as "UOP-SE."

Siamese CNN + Optical+ PolSAR: We used the VGG16 network to extract the features of the optical and PolSAR images and fuse such features by concatenation. We refer to this method as "SOP."

Siamese CNN + Optical+ PolSAR+SE": The SE block [22] was embedded into SOP after the concatenation of the encoded features. We refer to this method as "SOP-SE."

We designed six comparison methods to verify the performance of SDCAFNet. In summary, there are three comparison focuses. UNet-OP and UNet-PS are used to compare the performance difference between single-modal image input and multimodal image input. UOP and SOP are used to compare the performances of different feature extraction methods with SDCAFNet. UOP-SE and SOP-SE are used to compare the performances of different attention mechanism modules.

C. Hyperparameter Determination

We find that two hyperparameters have a significant impact on the segmentation result of SDCAFNet: the input image size H×W and the feature compression ratio t in SDCAM. SDCAFNet is an end-to-end semantic segmentation network, so the processing unit is an original image rather than an image patch. The size of the input image H×W directly affects the ability of SDCAFNet to extract spatial multiscale features. When H×W is too small, the small-scale features are fully extracted. Although subtle boundary segmentation can be more accurate, the trained model is not sensitive to large-scale global features, leading to oversegmentation that divides a wide range of continuous land covers into multiple subregions. If H×W is too large, the model lacks the ability to describe the details, and the segmentation

TABLE III
EXPERIMENTAL RESULTS OF THE DIFFERENT METHODS ON THE BAIYANGDIAN DATASET

Comparison methods	PA							OA	MIoU	FWIoU
	W	Bu	T	F	G	Ba	Un			
UNet-Op	75.89	50.77	74.08	71.76	71.32	54.31	27.61	63.3	52.23	58.06
UNet-PS	83.15	63.88	69.42	70.65	22.83	61.38	41.41	61.84	48.96	55.64
UOP	73.22	75.16	71.71	74.26	62.56	69.19	45.34	74.9	58.73	62.54
UOP-SE	77.59	**79.93**	74.6	**79.6**	64.91	**76.1**8	48.06	77.16	59.85	58.52
SOP	77.07	68.94	74	73.05	64.23	75.09	47.68	78.72	58.04	60.14
SOP-SE	93.9	77.79	78.43	73.56	**66.72**	62.51	**63.23**	77.32	61.71	60.97
SDCAFNet	**96.05**	74.44	**81.17**	76.02	65.22	68	62.35	**79.25**	**63.38**	**65.76**

The bold values denote the maximum of the columns.

TABLE IV
EXPERIMENTAL RESULTS OF THE DIFFERENT METHODS ON THE RONG DATASET

Comparison	PA							OA	MIoU	FWIoU
methods	W	Bu	T	F	G	Ba	Un			
UNet-Op	72.91	51.19	73.24	70.85	62.12	63.68	29.41	72.05	49.95	55.79
UNet-PS	89.08	66.08	70.09	70.22	24.89	69.49	39.32	70.1	54.55	55.84
UOP	72.57	74.52	71.87	71.22	61.26	70.4	48.15	71.05	56.75	59.72
UOP-SE	78.55	**77.63**	72.17	71.37	64.67	**75.69**	49.08	73.37	60.59	58.19
SOP	77.88	71.38	72.74	74.87	64.33	74.55	47.83	72.31	60.01	60.46
SOP-SE	91.07	72.61	**80.62**	75.28	65.25	63.84	54.57	74.49	60.42	62.77
SDCAFNet	**93.13**	72.16	77.92	**76.41**	**65.33**	65.07	**60.36**	**75.85**	**62.07**	**63.34**

The bold values denote the maximum of the columns.

performance of complex boundaries is too rough. The compression ratio t can affect the SDCAM spatial fineness for describing the channel relationship in the local region. When the input size **H**×**W** is fixed, if *t* is too large, there are too many parameters in the SDCAM, and the training process has difficulty converging. If *t* is too small, the description fineness of the local channel relationship is too rough, and the improvement in the feature fusion on the segmentation task is not obvious.

Therefore, in the experiment, H×W and *t* need to be adjusted together to find the optimal hyperparameter pair $<$ H×W and $t >$ that can realize the feature extraction ability of a reasonable spatial scale on the premise of rapid model convergence. Through many experiments, we found that when the number of hyperparametric pairs is $< 512 \times 512$ and $4 >$, SD-CAM has more reasonable parameters, SDCAFNet has a faster convergence speed, and the segmentation performance is greatly improved after feature fusion.

D. *Classification Over Two Datasets*

We evaluate the experimental results with several indexes, including the per-class accuracy (PA), overall accuracy (OA), mean intersection over union (MIoU), and frequency weighted intersection over union (FWIoU), in this article. To express the mathematical formulas of these evaluation metrics, we assume p_{ij} is the number of pixels of class *i* predicted to class *j* and T_i is the total number of pixels labeled to class *i*. *k* is the total number of classes. s_{ij} is a member of the confusion matrix *S*, and *C* is the total number of classes. Thus, the accuracy metrics are defined as follows.

PA: The simply computed ratio between the number of properly classified pixels and the total number of pixels for each class

$$\text{PA} = \frac{p_{ii}}{T_i} \tag{7}$$

where $i = 1, 2 \ldots, k$.

OA: The percentage of properly classified pixels and the total number of pixels in the entire image

$$\text{OA} = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k} T_i}. \tag{8}$$

MIoU: The mean percentage of the similarity between the prediction results and the ground truth

$$\text{MIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{s_{ii}}{\sum_{j=1}^{C} s_{ij} + \sum_{j=1}^{C} s_{ji} - s_{ii}}. \tag{9}$$

FWIoU: The standard metric to measure the similarity between the prediction results and the ground truth

$$\text{FWIoU} = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}}\sum_{i=1}^{N}\frac{\sum_{j=1}^{N}(s_{ij}s_{ii})}{\sum_{j=1}^{N}(s_{ij}+s_{ji}) - s_{ii}}. \tag{10}$$

1) Classification Results on the Baiyangdian Dataset: To verify the performance of the proposed method, we use six methods for comparison with SDCAFNet on the Baiyangdian dataset. The detailed experimental results are shown in Table III. Compared with UNet-OP and Unet-PS, we found that the optical image has higher average indexes, such as OA, MioU, and FWIoU, than those of the PolSAR image, which can preliminarily indicate that the optical image is more suitable for the mixed classification of multiple land-cover classes. However, the PolSAR image has a better classification performance for

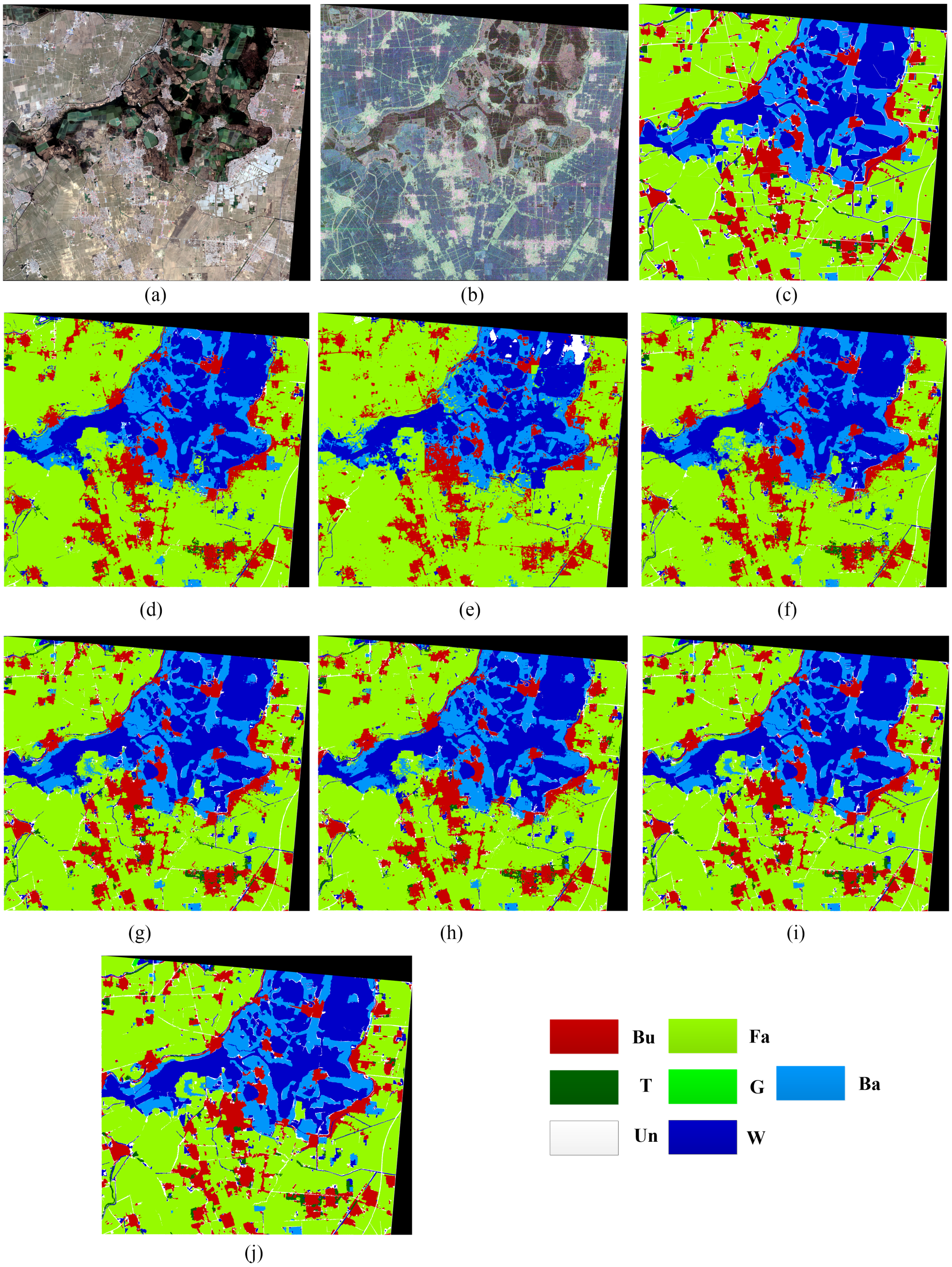


Fig. 7. Classification results of the different methods on the Baiyangdian dataset. (a) Optical image. (b) PolSAR (Pauli-RGB) image. (c) Ground truth. (d) UNet-OP. (e) UNet-PS. (f) UOP. (g) UOP-SE. (h) SOP. (i) SOP-SE. (j) SDCAFNet.

some single categories, such as water. Overall, all the methods for fusing optical images and PolSAR images improved all the indexes compared with the methods of single-modal images, showing that there is information complementarity between optical images and PolSAR images that is useful in the classification task. Comparing UOP and SOP, it can be observed that using the siamese network to extract optical and PolSAR features has better experimental performance than simple image concatenation at the input. Comparing UOP and SOP to UOP-SE and SOP-SE, it can be observed that channel selection with an attention module in the network can better solve the high-dimensional space problem of feature fusion. Further comparing SDCAFNet and SOP-SE, SDCAFNet has better experimental results, which shows that the local dense channel relationship captured by SDCAM is more suitable for land-cover semantic segmentation than the global channel relationship captured by SENet.

In conclusion, SDCAFNet has better experimental results than other methods. The reasons are as follows.

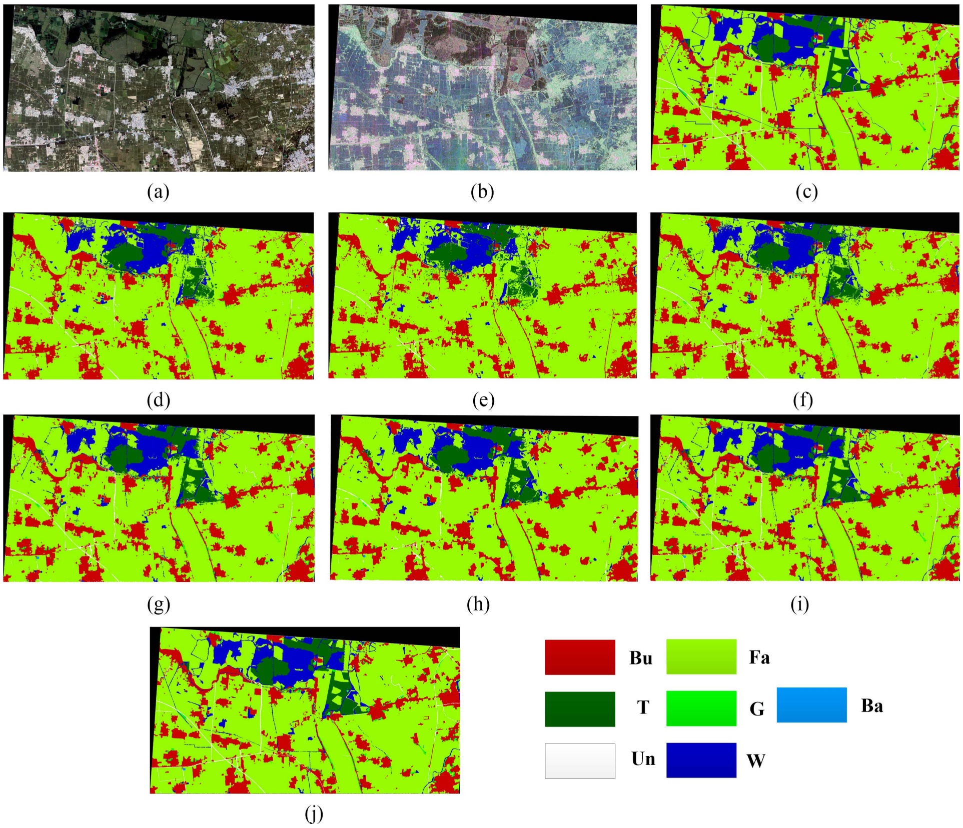


Fig. 8. Classification results of the different methods on the Rong dataset. (a) Optical image. (b) PolSAR (Pauli-RGB) image. (c) Ground truth. (d) UNet-OP. (e) UNet-PS. (f) UOP. (g) UOP-SE. (h) SOP. (i) SOP-SE. (j) SDCAFNet.

Compared with the input of a single-modal image, the multimodal input of optical and PolSAR images introduces more useful and complementary information.

1) The structure and pretraining method of the two-stream siamese convolution feature encoder can better capture the respective optical and PolSAR features.
2) Compared with other attention modules, the SDCAM has a better ability to capture dense and effective channel relationships.
3) Compared with the decoders in other methods, the symmetric skip-connection decoder block can recover spatiotemporal information and improve the resolution of the classification result graph to the greatest extent.

Fig. 7 shows the classification results of SDCAFNet and several other comparison methods. The results show that the classification results obtained by SDCAFNet are better than those of the other comparison methods, especially at the boundaries of different features. This proves that the fusion features extracted by SDCAM have a stronger ability to describe the detailed and dense relationships between local channels than other methods.

2) Classification Results on the Rong Dataset: The experimental results on the Rong dataset are shown in Table IV. It is clear that SDCAFNet has the best PA for water (W), farmland (F), grassland (G), and unknown (Un), while the OA, MIoU, and FWIoU are the best among all methods. Although the results for the Bu, trees (T), and bare land (Ba) are not optimal, they are not far from those of the optimal method. Overall, SDCAFNet shows the best classification effect in the Rong dataset upon comparison with the other methods. The classification effect of the abovementioned method is shown in Fig. 8.

It is clear from Fig. 8 that the SDCAFNet classification map is better than those of several comparison methods, demonstrating the superiority of this method. However, it is worth noting that grassland (G) and bare land (Ba) are easily confused.

E. Experimental Validation Curves

The experimental validation curves of the Baiyangdian and Rong datasets are shown in Figs. 9 and 10. Fig. 9 shows the process in which the loss gradually decreases and tends to stabilize with the increase in epochs during SDCAFNet training. Fig. 10 shows the process in which the accuracy gradually increases and tends to stabilize with the increase in epochs during SDCAFNet training. Because there is little difference in the final accuracies between the validation set and the training set, it can be observed that the training of SDCAFNet has not been overfitted and the trained model has good generalization.

F. Computational Time

The computational times for testing different methods per sample are shown in Table V. Experiments show that SDCAFNet can achieve better classification results on the premise of a computational time on the same order of magnitude.

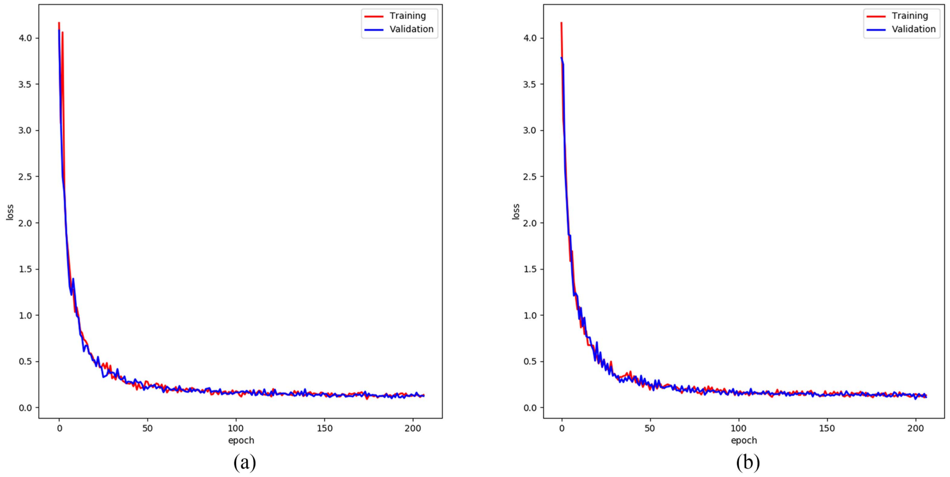


Fig. 9. SDCAFNet loss curves for the two datasets. (a) SDCAFNet loss curves on the Baiyangdian dataset. (b) SDCAFNet loss curves on the Rong dataset.

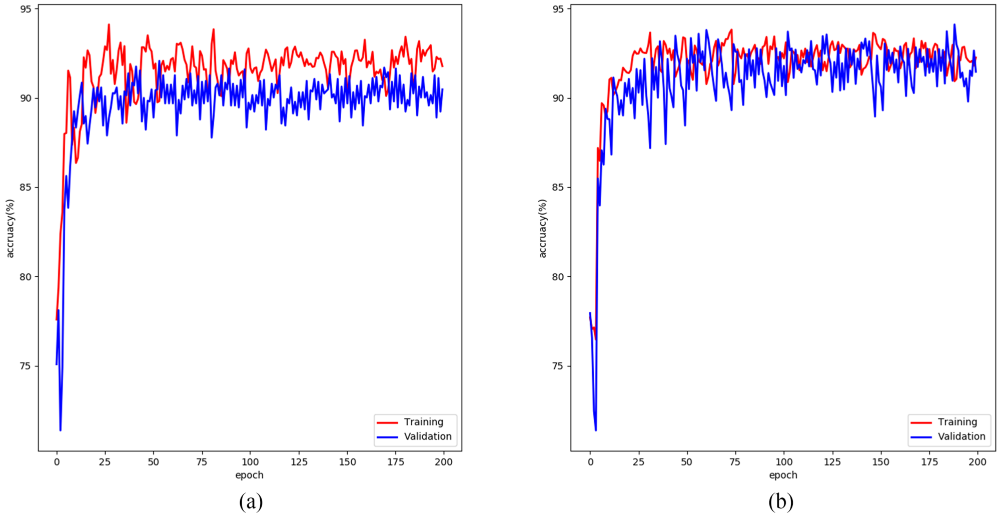


Fig. 10. SDCAFNet accuracy curves for the two datasets. (a) SDCAFNet accuracy curves on the Baiyangdian dataset. (b) SDCAFNet accuracy curves on the Rong dataset.

TABLE V
COMPUTATIONAL TIME OF DIFFERENT METHODS

Comparison methods	Computational time (sec.)/sample
UNet+Optical	3.15
UNet+PolSAR	3.21
UNet+Optical+ PolSAR	3.12
UNet+Optical+ PolSAR+SE	4.32
Siamese CNN + Optical+ PolSAR	4.10
Siamese CNN + Optical+ PolSAR+SE	4.12
SDCAFNet	4.34

V. Conclusion

In this article, a new end-to-end semantic segmentation network, SDCAFNet, was proposed based on optical and PolSAR image fusion. To capture dense and effective channel relationships, a new attention mechanism module, SDCAM, was proposed to extract the local relationships of multiple channels to improve the feature fusion performance of the network. Finally, to better recover the spatial-temporal information and improve the resolution of the segmentation result graph, we designed a symmetric skip-connection decoder block. Upon comparison with other methods on several indexes, it was shown that our method has higher accuracy and more practical value.

In the future, we hope to introduce the physical imaging mechanism of different sensors to the encoding process to optimize the performance of the multimodal network. In addition, we hope to study the research of the generation and transformation methods between optical and PolSAR images, which can enhance the training performance by solving the problem of insufficient co-registration data.

Acknowledgment

The authors would like to thank the support of the China Center for Resources Satellite Date and Application (CRESDA) with the optical and PolSAR remote sensing images used for this research.

REFERENCES

[1] H. Shi, W. Fu, and J. Huang, "Building segmentation in mountainous environment based on improved watershed algorithm," in *Proc. 3rd Int. Conf. Video Image Process.*, New York, NY, USA, 2019, pp. 168–172.

[2] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1633–1644, May 2018, doi: 10.1109/jstars.2018.2810320.

[3] H. Jing, Z. Wang, X. Sun, D. Xiao, and K. Fu, "PSRN: Polarimetric space reconstruction network for PolSAR image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10716–10732, 2021, doi: 10.1109/jstars.2021.3116062.

[4] X. Shi, S. Fu, J. Chen, F. Wang, and F. Xu, "Object-level semantic segmentation on the high-resolution Gaofen-3 FUSAR-map dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3107–3119, 2021, doi: 10.1109/jstars.2021.3063797.

[5] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021, doi: 10.1109/TGRS.2020.3015157.

[6] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 5518615, doi: 10.1109/TGRS.2021.3130716.

[7] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021, doi: 10.1109/TGRS.2020.3016820.

[8] D. Amarsaikhan, H. H. Blotevogel, J. L. van Genderen, M. Ganzorig, R. Gantuya, and B. Nergui, "Fusing high-resolution SAR and optical imagery for improved urban land cover study and classification," *Int. J. Image Data Fusion*, vol. 1, no. 1, pp. 83–97, Mar. 2010, doi: 10.1080/19479830903562041.

[9] Y. Ban, H. Hu, and I. M. Rangel, "Fusion of quickbird MS and RADARSAT SAR data for urban land-cover mapping: Object-based and knowledge-based approach," *Int. J. Remote Sens.*, vol. 31, no. 6, pp. 1391–1410, Mar. 2010, doi: 10.1080/01431160903475415.

[10] T. L. Ainsworth, D. L. Schuler, and J. S. Lee, "Polarimetric SAR characterization of man-made structures in urban areas using normalized circular-pol correlation coefficients," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2876–2885, Jun. 2008, doi: 10.1016/j.rse.2008.02.005.

[11] T. G. J. Rudner et al., "Multi 3 net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 702–709.

[12] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015, doi: 10.1109/JPROC.2015.2449668.

[13] M. J. Steinhausen, P. D. Wagner, B. Narasimhan, and B. Waske, "Combining sentinel-1 and sentinel-2 data for improved land use and land cover mapping of monsoon regions," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 73, pp. 595–604, Dec. 2018, doi: 10.1016/j.jag.2018.08.011.

[14] H. Zhang and R. Xu, "Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the pearl river delta," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 64, pp. 87–95, Feb. 2018, doi: 10.1016/j.jag.2017.08.013.

[15] H. Zhang, L. Wan, T. Wang, Y. Lin, H. Lin, and Z. Zheng, "Impervious surface estimation from optical and polarimetric SAR data using small-patched deep convolutional networks: A comparative study," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2374–2387, Jul. 2019, doi: 10.1109/jstars.2019.2915277.

[16] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4353–4361.

[17] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1011–1026, 2020, doi: 10.1109/jstars.2020.2975252.

[18] J. McGlinchy, B. Johnson, B. Muller, M. Joseph, and J. Diaz, "Application of UNet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3915–3918.

[19] D. Hong et al., "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020, doi: 10.1016/j.isprsjprs.2020.06.014.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2015, doi: 10.1109/tpami.2016.2572683.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham, Switzerland: Springer Int. Publishing, 2015, pp. 234–241.

[22] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representations*, Dec. 2014.

[23] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. 20th Int. Conf. Inf. Fusion*, Xi'an, China, 2017, pp. 1–7.

[24] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018, doi: 10.1109/lgrs.2018.2799232.

[25] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021, doi: 10.1109/tgrs.2020.2994057.

[26] 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, doi: 10.1109/TPAMI.2019.2913372.

[27] G. V. Laurin et al., "Optical and SAR sensor synergies for forest and land cover mapping in a tropical site in West Africa," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 21, pp. 7–16, Apr. 2013, doi: 10.1016/j.jag.2012.08.002.

[28] Z. Zhu, C. E. Woodcock, J. Rogan, and J. Kellndorfer, "Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using Landsat and SAR data," *Remote Sens. Environ.*, vol. 117, pp. 72–82, Feb. 2012, doi: 10.1016/j.rse.2011.07.020.

[29] V. Kumar, Y. S. Rao, A. Bhattacharya, and S. R. Cloude, "Classification assessment of real versus simulated compact and quad-pol modes of ALOS-2," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1497–1501, Sep. 2019, doi: 10.1109/lgrs.2019.2899268.

[30] M. Ghanbari, D. A. Clausi, L. Xu, and M. Jiang, "Contextual classification of sea-ice types using compact polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7476–7491, Oct. 2019, doi: 10.1109/tgrs.2019.2913796.

[31] J. Wang, B. Hou, L. Jiao, and S. Wang, "POL-SAR image classification based on modified stacked autoencoder network and data distribution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1678–1695, Mar. 2020, doi: 10.1109/tgrs.2019.2947633.

[32] D. Hong, Z. Pan, and X. Wu, "Improved differential box counting with multi-scale and multi-direction: A new palmprint recognition method," *Optik*, vol. 125, no. 15, pp. 4154–4160, Aug. 2014, doi: 10.1016/j.ijleo.2014.01.093.

[33] R. D. West, D. A. Yocky, B. J. Redman, J. D. Van Der Laan, and D. Z. Anderson, "Optical and polarimetric SAR data fusion terrain classification using probabilistic feature fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2097–2100.

[34] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, "FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data," in *Proc. Joint Urban Remote Sens. Event*, Dubai, United Arab Emirates, 2017, pp. 1–4.

[35] Y. Nan and W. Xi, "Classification of press plate image based on attention mechanism," in *Proc. 2nd Int. Conf. Saf. Produce Informatization*, Chongqing, China, 2019, pp. 129–132.

[36] Z. Yang, T. Zhang, and J. Yang, "Research on classification algorithms for attention mechanism," in *Proc. 19th Int. Symp. Distrib. Comput. Appl. Bus. Eng. Sci.*, Xuzhou, China, 2020, pp. 194–197.

[37] Y. Zhang and Z. Rao, "Hierarchical attention networks for grid text classification," in *Proc. IEEE Int. Conf. Inf. Technol. Big Data Artif. Intell.*, 2020, pp. 491–494.

[38] W. Lu, Y. Duan, and Y. Song, "Self-attention-based convolutional neural networks for sentence classification," in *Proc. IEEE 6th Int. Conf. Comput. Commun.*, 2020, pp. 2065–2069.

[39] L. C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[40] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6450–6458.

[41] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M^{3}Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4939–4949, Dec. 2018, doi: 10.1109/jstars.2018.2876357.

[42] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.

[43] J. Deng, W. Dong, R. Socher, L. J. Li, L. Kai, and F. F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[44] H. Sak et al., "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4280–4284.

[45] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Empirical Methods Natural Lang. Process.*, 2014.

[46] J. Koutník et al., "A clockwork RNN," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1863–1871.

[47] D. Quan et al., "Deep generative matching network for optical and SAR image registration," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6215–6218, doi: 10.1109/IGARSS.2018.8518653.

[48] M. Abadi, A. Agarwal, and P. Barham, "Tensorflow: Largescale machine learning on heterogeneous distributed systems," 2016. [Online]. Available: https://arxiv.org/abs/1603.04467

Boce Chu was born in Xingtai, Hebei, China, in 1991. He received the master's degree in information and communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016. He is currently working toward the Ph.D. degree majoring in information and signal processing with Beihang University, Beijing, China.

His present research interests focus on algorithms for remote sensing image processing based on artificial intelligence.

Jinyong Chen was born in Handan, Hebei, China, in 1970. He received the master's degree in communication and electronic system from the Institute of Communication Measurement and Control Technology, Shijiazhuang, China, in 1995.

He is currently a Research Scientist with the Key Laboratory of Aerospace Information Applications, China Electronics Technology Group Corporation, Beijing, China. He is the author of more than 20 technical papers. His present research interests focus on intelligent application of aerospace information.

Jie Chen (Senior Member, IEEE) was born in Zhengzhou, Henan, China, in 1973. He received the B.S. and Ph.D. degrees in information and communication engineering from Beihang University, Beijing, China, in 1996 and 2002, respectively.

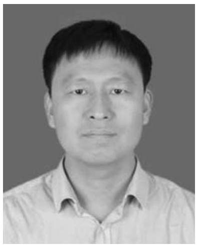
Since 2004, he has been an Associate Professor with the School of Electronics and Information Engineering, Beihang University. From 2009 to 2010, he was a Visiting Researcher with the School of Mathematics and Statistics, University of Sheffield, Sheffield, U.K., where he was involved in ionospheric effects on low-frequency space radars that measure forest biomass and ionospheric electron densities. Since 2011, he has been a Professor with the School of Electronics and Information Engineering, Beihang University. His research interests include multimodal remote-sensing data fusion, topside ionosphere exploration with spaceborne HF/VHF-SAR system, high-resolution spaceborne SAR image formation, and SAR image quality enhancement.

Xinyu Pei was born in Eerduosi, Inner Mongolia, China, in 1994. He received the master's degree in geographic cartography and geographic information engineering from Central South University, Changsha, China, in 2020.

His present research interests focus on geospatial pattern recognition.

Wei Yang (Member, IEEE) was born in Hubei, China, in 1983. He received the M.S. and Ph.D. degrees in signal and information processing from Beihang University (BUAA), Beijing, China, in 2008 and 2011, respectively.

In 2005, he studied the inner calibration signal analysis in synthetic aperture radar (SAR) systems. From 2006 to 2010, he focused on the system performance analysis and signal processing of highresolution and wide-swath mode in spaceborne SAR, including the multichannel terrain observation with progressive scan (TOPS) mode and the ScanSAR mode. Since 2011, he has been a Postdoctoral Researcher with the School of Electronics and Information Engineering, BUAA. His current research interests include ultrahigh-resolution spaceborne SAR image formation, modeling and data simulation, and novel techniques for spaceborne SAR systems.



Feng Gao was born in Handan, Hebei, China, in 1979. He received the master's degree in cartography and geographic information science from Wuhan University, Wuhan, China, in 2009. He is currently working toward the Ph.D. degree majoring in information and signal processing with Xidian University, Xi'an, China.

His present research interests focus on intelligent application of aerospace information.

Shicheng Wang was born in Xingtai, Hebei, China, in 1976. He received the master's degree in information and communication engineering from Xidian University, Xi'an, China, in 2009.

His present research interests focus on algorithms for remote sensing image processing based on artificial intelligence.