PolSAR Image Classification Via a Multigranularity Hybrid CNN-ViT Model With External Tokens and Cross-Attention

Wenke Wang, Jianlong Wang, Dou Quan, *Member, IEEE*, Meijuan Yang, *Member, IEEE*, Junding Sun, and Bibo Lu

***Abstract*—With the development of deep learning technology, the application of convolutional neural network (CNN) and vision transformer (ViT) for polarimetric synthetic aperture radar (PolSAR) image classification has been deepened. However, the PolSAR image has very rich information due to its special data form, which makes it difficult for the existing single network structure to comprehensively extract such effective information. In addition, deep learning methods require a large amount of data for training, whereas PolSAR labeled data is scarce and difficult to obtain. Therefore, a multigranularity hybrid CNN-ViT model based on external tokens and cross-attention is proposed for PolSAR image classification. First of all, CNN is able to learn local features very well. Thus, a CNN-based external feature extractor is designed to extract local features from the PolSAR image. Then, ViT can focus on global features. So, a multigranularity attention structure is constructed for extracting global information at multiple scales. With these two modules, the model can fully access the feature information contained in PolSAR images, which is more advantageous than a single network structure. Next, to further utilize these features, a cross-attention feature fusion module is built for fusing global–local information of different granularities. Finally, by connecting with the softmax classifier, the network outputs the final prediction results. Experimental results on three benchmark datasets show that the present method using a small amount of labeled data for training also achieves the highest level of classification among the compared methods.**

***Index Terms*—Convolutional neural network (CNN), cross-attention, external tokens (ETs), multigranularity, polarimetric synthetic aperture radar (PolSAR), vision transformer (ViT).**

I. Introduction

SYNTHETIC aperture radar (SAR) [1] is an active microwave imaging method. It provides its own illumination and can therefore operate around the clock without the influence of the sun. In addition, it utilizes microwaves to penetrate clouds and tree canopies, soil and snow. So, it also has the ability to deal with different weather conditions [2], [3]. Thanks to these characteristics, SAR plays an important role in all aspects of national production and life, such as land surveying [4], planning [5], and utilization in urban and rural areas [6], etc. [7]. Polarimetric synthetic aperture radar (PolSAR) is a type of SAR that can operate in different polarization combination modes. Due to the variety of polarization combinations, it is possible to obtain richer information data. Comparatively, the exploitation of PolSAR allows a new level of information acquisition for different kinds of land cover types [8]. PolSAR image classification [9] is the process of classifying all pixels in an image into a certain category according to specific rules. It is a key fundamental research direction of PolSAR image understanding and interpreting technology, as well as a prerequisite work for feature recognition and evaluation of images. However, the special structural form of PolSAR data [10] allows it to contain richer information. As a result, it is difficult for most deep learning methods to fully learn the feature information of PolSAR data [11]. At the same time, PolSAR labeled data are scarce and relies heavily on expert knowledge as well as manual manipulation. While, deep learning methods basically sample a large amount of labeled data in order to perform well. Therefore, these issues make the application of deep learning methods to PolSAR image classification a challenging topic.

Before deep learning methods were applied to the PolSAR image classification task, a number of different approaches have been given. The more traditional methods are based on polarimetric targets decomposition. The methods decompose the obtained polarimetric data into some parameters of practical physical significance, which facilitates the analysis of the complex scattering processes of the targets [12], [13], [14], [15]. In addition, the methods based on statistical distribution are continuously provided for PolSAR image classification [16], [17], [18]. For example, in view of the problem of class distribution bias in datasets between different domains, an unsupervised domain adaptive network based on coordinated attention and weighted clustering was presented for achieving the alignment of data distributions between different domains [19]. In the last decade, the use of machine learning methods for PolSAR image classification tasks has become a major trend [20], [21], [22].

Manuscript received 22 January 2024; accepted 29 March 2024. Date of publication 3 April 2024; date of current version 16 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62201201, Grant 62201407, and Grant 62301443, in part by the Doctoral Foundation of Henan Polytechnic University under Grant B2022-15, and in part by China Postdoctoral Science Foundation under Grant 2022M722496. *(Corresponding author: Jianlong Wang.)*

Wenke Wang, Jianlong Wang, Junding Sun, and Bibo Lu are with the School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, China (e-mail: wangjianlong24@hpu.edu.cn).

Dou Quan is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: quandou@xidian.edu.cn).

Meijuan Yang is with the School of Artificial Intelligence OPtics and ElectroNics, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: mjuanyang@nwpu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3384420

Meanwhile, the link between machine learning and statistical data distribution methods is also becoming a topic in research. Such as Wang et al. [23] presented an improved autoencoder for PolSAR image classification. It combines polarimetric targets decomposition parameters selection as well as statistical distribution of PolSAR data to improve the autoencoder network for better classification. Furthermore, there was a work that combined Wishart measures into machine learning training, which led to the definition of novel Wishart-AE and Wishart-CAE models [24]. In recent years, accompanied by the persistent research of deep learning methods, the combinations of traditional methods have erupted into a new research fervor, which has injected new energy into the research of PolSAR image classification.

Benefiting from the excellent hierarchical and local feature extraction ability of convolutional neural networks (CNNs), many improved networks [25], [26], [27] have been designed to mine more abstract and deeper features in PolSAR images. Wang et al. [28] considered that most CNN-based methods can only classify one pixel at a time, thus ignoring the inherent correlation between different feature types. So, a type of CNN using a fixed feature size was constructed to classify all pixels in a patch at once. In order to fully utilize the phase information of PolSAR images, Li et al. [29] redefined the regular operations of CNN in the complex domain and represented the data through complex-valued matrices, which in turn led to the proposal of a complex multiscale network. To achieve PolSAR image classification using a deep learning approach on a limited labeled dataset, Shang et al. [30] presented a spatial feature-based CNN. By using a two-branch structure and sharing parameters, the network is able to receive more than one sample as input. The original dataset can be expanded by combining different samples through such a special structure, which relieves the problem of insufficient training data in the PolSAR image classification task. Parikh et al. [31] pointed out that fewer studies have explored the effect of convolutional kernel size selection on classification modeling. Hence, a CNN based on homogeneous kernel selection was introduced for PolSAR classification modeling. The method demonstrates that classification accuracy can be effectively improved by using a pair of CNN models with different kernel sizes to classify image blocks with different high and low homogeneity separately. Aiming at the problem that buildings in PolSAR images are easily confused with other objects to affect the feature extraction, Li et al. [32] constructed a method to extract buildings from PolSAR images with statistical texture, polarization features, and hyperpixels by constructing a feature space that is sensitive to buildings. Although CNNs can learn abstract features in PolSAR images well, each convolutional kernel is limited to a fixed and small region, making it difficult to effectively extract global correlations in PolSAR images.

In recent years, with the wide applications of vision transformer (ViT) [33] in the field of vision [34], [35], [36], many researchers have tried to introduce it into the field of remote sensing. Liu et al. [37] designed an end-to-end network for HR SAR classification, i.e., global–local network structure. The structure employs a lightweight CNN and a compactly structured ViT to learn local and global features, respectively. Moreover, the complementary information is mined through the fusion network, and a better classification effect is obtained. Liu et al. [38] established a new lightweight attention discarding transformer in order to solve the problems of small training samples, easy overfitting and lack of local information extraction ability in the application of ViT for HR SAR. The model is based on the swin transformer backbone network, which introduces lightweight group convolution and channel shuffling block instead of the self-attention mechanism to extract local features, and consequently a new method of composite normalization is presented. Dong et al. [39] first explored the application of ViT for PolSAR image classification and designed a ViT-based representation learning framework. Through dividing PolSAR images into small patches and converting them into token vectors that can be accepted by the encoder, the framework realizes the extraction of global features of PolSAR images and achieves good classification results. Yin et al. [40] introduced ViT to the study of the classification of multitemporal PolSAR data. A classification model is devised that combines a dual-stream network and a temporal–polarimetric–spatial transformer for extracting temporal–polarization–spatial features. Moreover, a 3-D convolutional attention module is developed to weigh the importance of different dimensions. Li et al. [41] introduced a multifeature dual-stage cross manifold attention network, which improves the feature extraction capability of the network by mining the complementary information between different features. The network presents a cross-feature network module to acquire different feature information in PolSAR targets in the first stage and a cross-metamorphic attention transformer to extract nonlinear relationships between features in the second stage. Despite the fact that ViT has been widely researched and applied in the field of remote sensing, ViT-based PolSAR image classification methods are still scarce compared to CNN-based methods.

Considering the better utilization of the local feature extraction capability of CNN and the global information acquisition capability of ViT to solve the problem of incomplete feature extraction from PolSAR data, this article proposes a multigranularity hybrid CNN-ViT model for PolSAR image classification based on external tokens (ETs) and cross-attention. The model requires only a very small amount of labeled data to complete the training and obtain more effective feature information from PolSAR data. Compared to the works that already exist, the main contributions of this article are as follows.

1) Without changing the structure of the ViT, an ET module is proposed to extract the local information in the PolSAR image. Then, the local information is converted into a feature vector form that can be accepted by the ViT encoder. Such an approach provides ViT with the ability to access local information without compromising the global nature of ViT and is highly scalable. Furthermore, a multigranularity attention (MGA) structure is constructed for extracting global information at different granularities. The use of both modules can fully extract the rich feature information contained in PolSAR data and solve the problem that it is difficult for a single network structure to fully learn the information of PolSAR data.

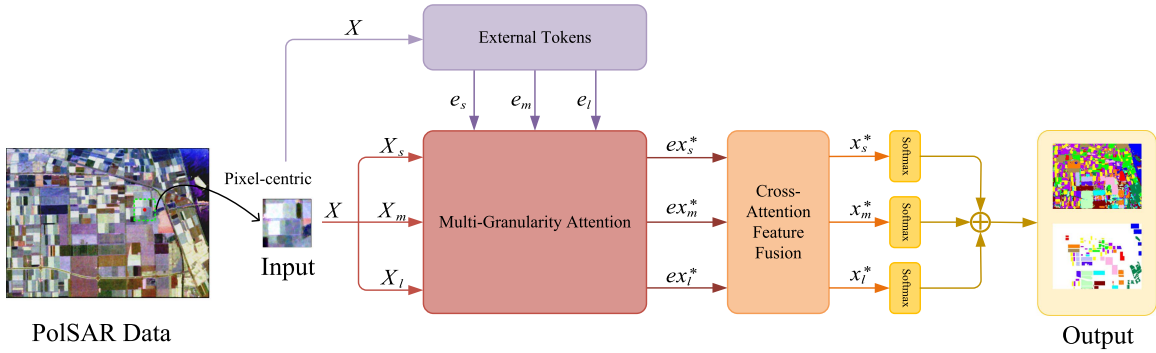


Fig. 1. Overall structure of the proposed model.

2) To further utilize the ET module and the MGA structure to obtain global–local information, a cross-attention feature fusion (CAFF) module is proposed for complementary fusion of features extracted from different granularity branches. It fully explores the feature extraction capabilities of CNN and ViT. By fusing global and local feature information, the characterization representation capability of the model is further enhanced. In this way, the extraction of complex features from PolSAR data can be accomplished with only a very small amount of labeled data. As a result, the conflict between the need for large amounts of labeled data for current deep learning methods and the scarcity and difficulty of accessing PolSAR labeled data is resolved.
3) To better realize the use of CNN in conjunction with ViT, this article further discusses the effect of local feature incorporation on the global performance of ViT. With a special experimental design and the introduction of the local impact index (LII) as a metric, the question of how CNN and ViT can be effectively combined to better extract features from PolSAR data is fully explored. The combined use of multiple network structures will be an interesting direction for PolSAR image classification.

The rest of this article is organized as follows. Section II describes the model structure and related principles. Section III details the PolSAR datasets used for the experiments, the experimental parameters, and the analysis of the experimental results. Section IV provides a discussion of the different factors that affect model performance. Finally, Section V concludes this article.

II. Proposed Methods

The classification performance of PolSAR images depends on the extraction of learned features from the complex polarized spatial features of PolSAR images. Therefore, a multigranularity hybrid CNN-ViT model based on ETs and cross-attention is presented for PolSAR image classification. The structure of the proposed method is composed of three parts, as shown in Fig. 1. The first part is the ET module, which is used to extract local features from PolSAR images. The second part is a MGA architecture for extracting global features with different granularity sizes. The third part is the CAFF module for the complementary fusion of global–local features of different granularities. After feature extraction, the classification task is performed on each of the three granularity branches and the sum of the predicted probabilities of the three branches is taken as the final classification result. The algorithmic flow of the proposed method is summarized as Algorithm 1.

Before performing PolSAR image feature extraction, some processing on the raw PolSAR data are conducted to make it conform to the input data structure of CNN and ViT. To begin with, the initial form of the PolSAR data is a 9-D real vector [42]: $[T_{11}, T_{22}, T_{33}, \mathrm{Re}[T_{12}], \mathrm{Im}[T_{12}], \mathrm{Re}[T_{13}], \mathrm{Im}[T_{13}], \mathrm{Re}[T_{23}], \mathrm{Im}[T_{23}]]$, where $\mathrm{Re}[\bullet]$ and $\mathrm{Im}[\bullet]$ denote the real and imaginary parts, respectively. The initial form of the PolSAR data can then be expressed as $\mathbb{R}^{9\times h\times w}$, where h and w denote the height and width of the PolSAR image, respectively. After that, a domain of size $p \times p$ is extracted for each pixel point contained in the image, where the edge portion is subjected to a complementary zero strategy to ensure the completeness of the extracted domain. For each domain can be represented as $\mathbb{R}^{9\times p\times p}$, as a result the input data for feature extraction can be represented as $\mathbb{R}^{(h\times w)\times 9\times p\times p}$.

A. External Tokens

Since CNNs have strong local feature extraction capability and have been employed as feature extractors for PolSAR images in many works [43], [44], [45], [46], a multilayer CNN is designed for extracting local features from PolSAR images. The CNN involves three blocks consisting of convolutional and pooling layers. Each has a convolutional kernel size of 3. ReLU is utilized as an activation function in each block to enhance its nonlinear representation. To use the feature maps extracted by the CNN directly as ETs for ViT, the number of convolutional channels in each block of the CNN is adapted. Assume that the three granularity branches are coarse-grained L, medium-grained M, and fine-grained S, and the corresponding granularity sizes g of each branch are l, m, and s, respectively. As shown in Fig. 2, $X_{(i,g)}$ denotes the output of each convolutional block, where $(i, g) \in \{(1, s), (2, m), (3, l)\}$ denotes the correspondence between different levels of external feature tokens and different granularity of branching. Then, $X_{(i,g)}$ can

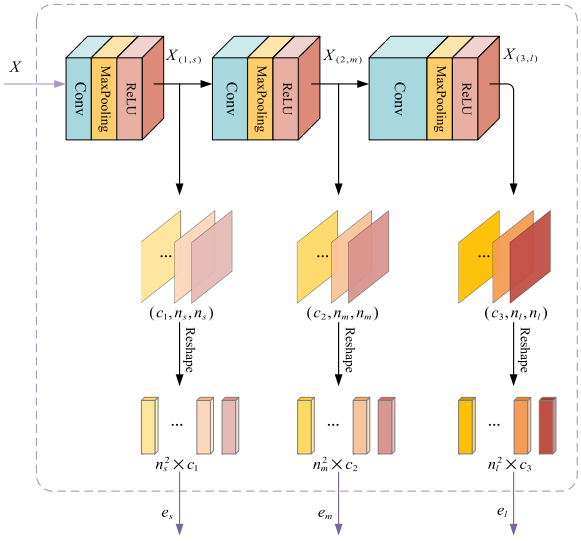


Fig. 2. Structure of the ET module.

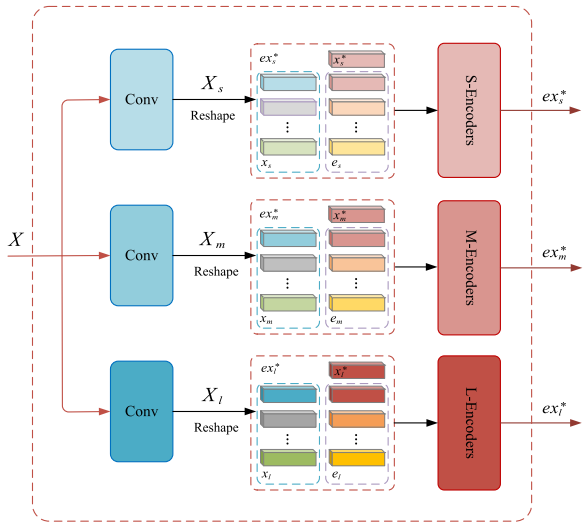


Fig. 3. Structure of the MGA module.

be obtained by the following equation:

$$X_{(i,g)} = \text{ReLU}(\text{Pooling}_i(\text{Conv}_i(X_{(i-1,g)}))) \tag{1}$$

where Conv_i and Pooling_i denote the convolutional and pooling layers of the ith block, respectively. When i is 1, $X_{(i-1,g)}$ is the input data X. First for the input data $X \in \mathbb{R}^{9\times p\times p}$, certain underlying features $X_{(1,s)}$ with shape (c_1, n_s, n_s) are extracted after a convolution pooling operation in the first convolution block. A second block of $X_{(1,s)}$ is then performed on $X_{(2,m)}$ of shape (c_2, n_m, n_m). Finally after the third block the output shape $X_{(3,l)}$ with (c_3, n_l, n_l) is obtained. The feature maps $X_{(1,s)}$, $X_{(2,m)}$, and $X_{(3,l)}$ obtained after the feature extractor will be provided as output. In addition, the three feature maps are deformed at output time for direct use as ETs. $X_{(i,g)} \in \mathbb{R}^{c_i\times n_g\times n_g}$ will be deformed from 3-D feature maps to 2-D feature vectors $e_{(i,g)} \in \mathbb{R}^{n_g^2\times c_i}$. To guarantee that the feature vectors have the same length as the token vectors in the multigranularity branch, the values of n_g and c_i need to satisfy the following conditions:

$$n_g = \lfloor \frac{p}{g} \rfloor \quad c_i = 9 \cdot g \cdot g \tag{2}$$

where n_g denotes the shape of the output of the convolutional layer corresponding to the g granularity branch, and its value is rounded down; c_i stands for the dimension of the convolutional output of the ith layer. Ultimately, the ETs can be represented as $e_g \in \mathbb{R}^{n_g^2\times(9\cdot g^2)}$.

The module can output feature maps of different sizes with different receptive fields, and can be added to the ViT branch of the corresponding granularity as a complement to the original patch tokens. It effectively makes up for the lack of local feature extraction capability of ViT, which enables it to obtain more comprehensive and rich feature information. Furthermore, if the input data to the feature extractor are changed to other representations of the PolSAR image, or other data modalities, it will give the proposed method the capability of multimodal data processing and feature extraction. Therefore, the ET is very flexible and scalable, which greatly enhances the feature extraction and representation learning capabilities of the proposed method.

B. Multigranularity Attention

Benefiting from the excellent global information extraction capability of ViT, it will be applied as the backbone network to extract the global features of PolSAR images [47], [48], [49]. The structure of the MGA module is shown in Fig. 3. In order to extract global features with different granularities, three parallel ViT structures with changes made at the data processing layer are set up first. For the input data $X \in \mathbb{R}^{9\times p\times p}$ with the same three branches, it is first divided into different granularity patches $X_g \in \mathbb{R}^{n_g^2\times(9\times g\times g)}$, where $n_g = \lfloor \frac{p}{g} \rfloor, g \in \{l, m, s\}$ denotes the number of corresponding granularity patches. Subsequently, X_g are flattened to 2-D and the patch tokens $x_g \in \mathbb{R}^{n_g^2\times(9\cdot g^2)}$ are obtained. At this point, splice the ETs $e_g \in \mathbb{R}^{n_g^2\times(9\cdot g^2)}$ into the current token vectors to get new patch tokens $ex_g \in \mathbb{R}^{(n_g^2+n_g^2)\times(9\cdot g^2)}$. Then, a class token $x_g^* \in \mathbb{R}^{1\times(9\cdot g^2)}$ is added to each of the granularity branch token vectors for capturing the feature representation of that branch. Finally, the token vectors for global feature extraction can be represented as $ex_g^* \in \mathbb{R}^{(n_g^2+n_g^2+1)\times(9\cdot g^2)}$. Be aware that for a fixed input image size p it may not be possible to divide the patches equally. To accommodate this situation, this article improves on the original patch division method in ViT. A 2-D convolutional layer with a convolutional kernel of the same size as the step size is used in the data preprocessing layer of ViT instead of the original sliding window approach to implement the tokenization process. Positional coding has also been removed to make patch division more flexible and unrestricted.

Once the token vectors with different granularity sizes are captured, each branch extracts the global information contained in the token vectors via multiple encoders. Each encoder can be composed of multihead self-attention (MSA), multilayer perceptron (MLP), and layer normalization (LN), and residual

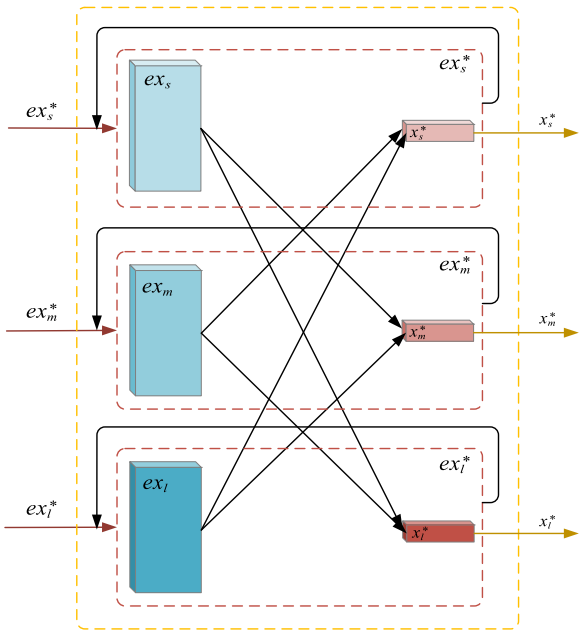


Fig. 4. Structure of the CAFF module.

connections are used between each layer. Therefore, the way the encoder works can be described by the following equations:

$$x'_n = \text{MSA}(\text{LN}(x_{n-1})) + x_n \tag{3}$$

$$x_n = \text{MLP}(\text{LN}(x'_n)) + x'_n, n \in \mathbb{R}^N \tag{4}$$

where n denotes the nth layer encoder and x_n denotes the data processed by the nth layer encoder. Following feature extraction by multiple encoders, the class tokens x_g^* of each of the three branches contain all the features represented by the token vectors in the branch.

Such MGA structure can obtain richer feature information, and the information contained for different granularity can be well extracted, which more fully utilizes the PolSAR image. At the same time, the addition of ETs also enables the structure to obtain more local information, which together with the original token vectors will enhance the robustness of the model and make the model more capable of feature representation.

C. Cross-Attention Feature Fusion

The two previous works are just obtaining a global–local feature representation of the PolSAR image at multiple granularities, but there is a lack of complementary fusion of information among the three mutually independent branches. Thus, the CAFF module is applied to the complementary fusion of the features obtained from the three branches, which consists of multihead cross-attention (MCA), LN, and residual connection. Because of the addition of class tokens, the abstract information contained in the patch tokens in each branch can be represented by the class tokens of that branch. So, feature fusion can be realized in the following way, as presented in Fig. 4.

Specifically, the class tokens of each branch act as agents for that branch, exchanging information with the patch tokens of the other branches, which are then reprojected to the original branch. Repeating this operation several times, the class tokens in each branch contain the information of the patch tokens in the other branches, which accomplishes the complementary fusion of different granularity information. When the token vectors of the three branches are fed into the CAFF module, in the case of the L branch, for example, the class token X_l^* is transformed by a linear projection $f_m(\cdot)$ into token vectors $f_m(x_l^*)$ of the same shape as the patch tokens ex_m of the M branch. Then, $f_m(x_l^*)$ pass through the projection matrix W_q to form the variable query q in the self-attention, while $f_m(x_l^*)$ and ex_m together form the new vector set $[f_m(x_l^*) \,||\, ex_m]$ and go through the projection matrix W_k and W_v into variable key k and variable value v, respectively. After the self-attention computation, the feature information in the M patch tokens ex_m are incorporated in $f_m(x_l^*)$, which are then retransformed to its original shape by a linear projection $f_l(\cdot)$ and spliced with the patch tokens ex_l to form new L token vectors. Afterward, in the same way, the information contained in the patch tokens ex_s for the branch S is fused in x_l^*. In this way, L completes feature fusion with the other two branches, and its class token contains the information in the patch tokens of the other two branches. Similarly, the class tokens in M and S will complete the information exchange with the patch tokens of the other two branches, respectively. Following the cross-attention computation, the feature information of the other two branches is fused in the class tokens of all three branches, which accomplishes the complementary fusion between different branches. The process can be expressed in the following equations:

$$x_g^{*\prime} = f_{\bar{g}}(x_g^*) + \text{MCA}(\text{LN}([f_{\bar{g}}(x_g^*) \,||\, ex_{\bar{g}}])) \tag{5}$$

$$x_g^* = [f_g(x_g^{*\prime}) \,||\, ex_g], g \in \{l, m, s\} \tag{6}$$

where g denotes branch of different granularity and $\bar{g}$ signifies other branches different from g. Since the variable q in MCA consists of only class token, it reduces a lot of computation in the actual attention computation, which makes it more efficient.

The module is able to effectively fuse the extracted feature information from multigranularity branches in a fully complementary manner. A softmax classifier is then connected separately to obtain the prediction results, and the sum of the prediction results of the three branches will be taken as the final classification result, which further improves the credibility of the prediction results of the model.

III. Experimental Results and Analysis

In this section, the information about the dataset used for the experiments is presented first and then the specific parameter configurations of the proposed model are illustrated. In addition, the evaluation indicators used for the experiments are described. In the end, the arguments based on the experimental results are analyzed.

A. Dataset

In order to better validate the effect of the proposed model, three different PolSAR standard datasets were selected, which

Algorithm 1: Multi-Granularity Hybrid CNN-ViT Model with External Tokens and Cross-Attention.

Input: The coherency matrix of PolSAR data and its corresponding ground truth; parameters of the model shown in Table II; the number of cross-validation folds; data batch size; training epochs; learning rate; weight decay

Output: Classification results

1: Extract the neighborhood by the labeled PolSAR data center pixel as the input image and form the corresponding dataset by using the center pixel label as the image label.
2: A specified number of randomly selected data from the labeled datasets are originally used as the training and validation sets, and the remaining portion as the test set.
3: The training data are fed into the proposed ET module to obtain three different levels of local features with Eq. (1).
4: Divide the training data into three patch sequences with different granularities, deform the local feature maps obtained earlier and add them to the corresponding granularity branches. Global information at different levels of granularity is obtained through the constructed multi-granularity attention structure with Eq. (3)–(4).
5: Three different granularities of feature information are further complementarily fused using the proposed cross-attention feature fusion module with Eq. (5)–(6).
6: Each granularity branch generates the corresponding predicted category probability by the softmax classifier separately, and the sum of the three branch probabilities is the final result.
7: The predicted image is obtained by classifying all the pixels with the trained model.

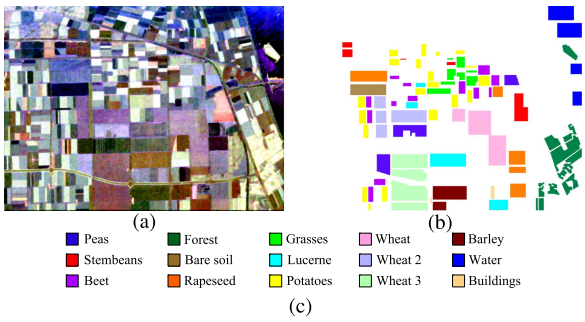


Fig. 5. AIRSAR Flevoland dataset and its color code. (a) Pauli-RGB image. (b) Ground truth map. (c) Legend of the dataset.

were acquired from different platforms. Figs. 5–7 show their Pauli-RGB image, ground truth map, and a legend illustrating the correspondence between each color and feature type. The labeling information of the pixels is manually performed by expert knowledge [50].

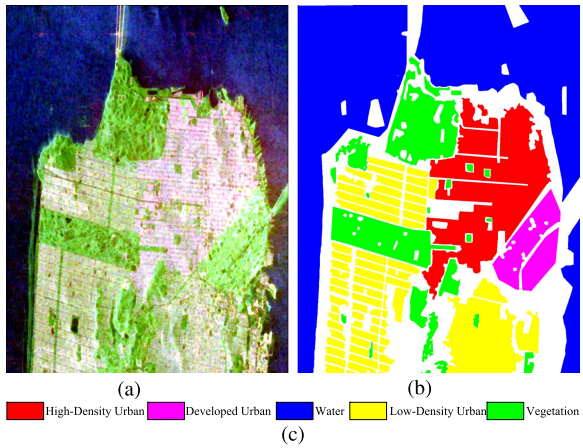


Fig. 6. RADARSAT-2 San Francisco dataset and its color code. (a) Pauli-RGB image. (b) Ground truth map. (c) Legend of the dataset.

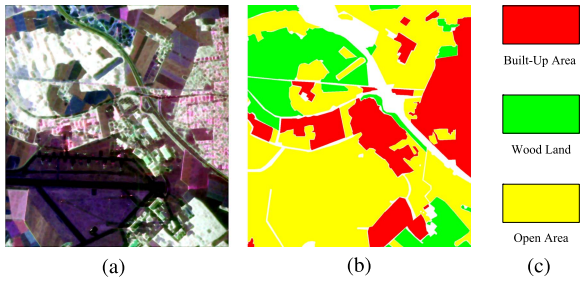


Fig. 7. ESAR Oberpfaffenhofen dataset and its color code. (a) Pauli-RGB image. (b) Ground truth map. (c) Legend of the dataset.

1) AIRSAR Flevoland: Fig. 5 illustrates a submap Flevoland image of size 750 × 1024 of the L-band multiview PolSAR dataset acquired by the AIRSAR platform. The dataset contains 15 feature categories with each category identified by a color. The total number of pixels labeled in the ground truth map is 167 712.

2) RADARSAT-2 San Francisco: Fig. 6 is a C-band dataset of the San Francisco area acquired by the RADARSAT-2 satellite with a selected scene size of 1380 × 1800. The dataset includes five feature types: High-density urban, water, vegetation, developed urban, and low-density urban. The total number of pixels that have been labeled in the ground truth map is 1 804 087 [51].

3) ESAR Oberpfaffenhofen: Information on the L-band PolSAR dataset of the Oberpfaffenhofen area with a size of 1300×1200 acquired by the ESAR airborne platform is presented in Fig. 7. Three categories of the dataset are built-up area, wood land, and open area. The total number of pixels with labeling information in the ground truth map is 1 374 198.

In conclusion, a summary of the specific information of the PolSAR datasets is presented in Table I.

B. *Parameter Configuration and Training Details*

The specific network configuration of the proposed method is as follows. Since ViT receives input in the form of imagery, the raw PolSAR data needs to be processed. For all the data

TABLE I
SPECIFIC INFORMATION OF THE POLSAR DATASETS

Dataset	Platform	Band	Year	Resolution	Size	Categories
Flevoland	AIRSAR	L	1989	6.6 × 12.1m	750 × 1024 pixels	15
San Francisco	RADARSAT-2	C	2008	10 × 5m	1800 × 1380 pixels	5
Oberpfaffenhofen	ESAR	L	2002	3.0 × 2.2m	1300 × 1200 pixels	3

TABLE II
SPECIFIC PARAMETER CONFIGURATIONS FOR THE THREE MODULES

Module	Parameter configuration		
ET	Block 1	Block 2	Block 3
	Conv.36@3 × 3 Maxpool@3 × 3	Conv.81@3 × 3 Maxpool@3 × 3	Conv.225@3 × 3 Maxpool@3 × 3
	(36,7,7)	(81,5,5)	(225,3,3)
MGA	S	M	L
	s=2 depth=1 MSA-(3 × 64),36 FFN-108	m=3 depth=1 MSA-(3 × 64),81 FFN-243	l=5 depth=1 MSA-(3 × 64),225 FFN-675
CAFF	depth=2 MSA-(3 × 64)		

used in this article, a neighborhood of size 14×14 is extracted centered on the pixel point, resulting in a value of 15 for the input space size p. This value is chosen for two reasons. One is inspired by Dong et al. [39]. To maintain consistency, this setting has been extended. Another one is that the size of the input space affects the amount of information in the input data and the calculation cost of the model. On the one hand, a larger neighborhood can contain more data information, but it will increase the computational cost of the model. On the other hand, a small neighborhood will reduce the computational overhead of the model, but the amount of information it contains will also be reduced. After the previous experimental verification, 15 is a more appropriate choice. The specific parameter configurations for the three modules can be viewed in Table II.

Different amounts of training data are used for each dataset in order for the model to be adequately trained. Three datasets of 300 (2.68%), 900 (0.25%), and 1500 (0.33%) randomly selected labeled data from each category are used for training and validation. Parentheses indicate the proportion of selected data to the total labeled data. All remaining labeled data will be available for testing. Selecting training data by category alleviates the sample imbalance problem to some extent, and allows categories with less data to be well represented by the model.

In addition, a five-fold cross-validation approach is utilized when performing the training process. Each fold has 50 training epochs and each batch size is 256. The Adam optimizer is also applied to automatically adjust the learning rate with an initial value set to 0.001. The loss function employs a cross-entropy loss. With model testing on all test data, the one-fold model with the highest OA among the five-fold models is selected as the final training model and used for subsequent prediction of results.

This article focuses on exploring the deep combination of CNN and ViT for better feature extraction. Therefore, CNN-based methods ResNet [52], CV-FCN [53], and CV-3D-CNN [54], ViT-based method SViT [39] and CNN combined with ViT methods CCT [55] and MCPT [56] are selected for comparison experiments.

C. Objective Evaluation Indicators of Classification Performance

The experiments use overall accuracy (OA) [57], average accuracy (AA), and Kappa coefficient (Kappa) [58] as the evaluation indicators to judge the classification performance of the model. OA is a measure of the OA performance of the model and can visualize the performance strengths and weaknesses of the model. The formula is as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{7}$$

where TP stands for true positive, FN denotes false negative, FP expresses false positive, and TN symbolizes true negative. AA is the average of the prediction accuracy for each feature type and can be applied to measure the classification performance of the model on a specific feature type. AA is calculated by the following formulas:

$$\mathrm{AA} = \frac{\sum \mathrm{Recall}_i}{N_i} \tag{8}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

where Recall is the ratio of the number of correctly categorized positive samples to the number of positive samples, i means the category, and N_i signifies the number of categories. Kappa is calculated based on the confusion matrix for consistency testing, where larger values indicate higher consistency and better model classification performance. Kappa can be calculated by the following formula:

$$\mathrm{Kappa} = \frac{p_o - p_e}{1 - p_e} \tag{10}$$

where p_o indicates the OA. With $a_1, a_2, \ldots, a_c$ pointing the number of true samples in each class, $b_1, b_2, \ldots, b_c$ showing the number of predicted samples in each class, and n signifying the total number of samples, p_e can be expressed as follows:

$$p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \cdots + a_c \times b_c}{n^2}. \tag{11}$$

From the above-mentioned calculation formulas, the three evaluation indicators can well measure the performance of the

TABLE III
OBJECTIVE EVALUATION INDICATORS OF SEVEN METHODS ON THE AIRSAR FLEVOLAND DATASET

	ResNet	CV-FCN	CV-3-D-CNN	SViT	CCT	MCPT	Proposed
Water	0.6812 ± 0.2065	0.9496 ± 0.0049	0.9866 ± 0.0036	0.9882 ± 0.0131	0.9822 ± 0.0304	0.9110 ± 0.0814	**0.9985 ± 0.0021**
Forest	0.9240 ± 0.0756	0.9940 ± 0.0015	**0.9990 ± 0.0006**	0.9787 ± 0.0074	0.9933 ± 0.0031	0.9791 ± 0.0121	0.9807 ± 0.0271
Lucerne	0.8362 ± 0.1876	0.8855 ± 0.0009	0.9868 ± 0.0045	0.9905 ± 0.0064	0.9895 ± 0.0061	0.9861 ± 0.0032	**0.9952 ± 0.0042**
Grass	0.7900 ± 0.2771	0.9013 ± 0.0036	0.9670 ± 0.0027	0.9828 ± 0.0171	0.9305 ± 0.0170	0.9715 ± 0.0049	**0.9893 ± 0.0122**
Peas	0.9824 ± 0.0185	0.9934 ± 0.0028	0.9963 ± 0.0019	0.9787 ± 0.0130	0.9973 ± 0.0019	0.9904 ± 0.0018	**0.9989 ± 0.0014**
Barley	0.5080 ± 0.4420	0.8420 ± 0.0053	0.8231 ± 0.0021	0.9935 ± 0.0108	0.9678 ± 0.0666	0.9749 ± 0.0165	**0.9983 ± 0.0014**
Bare Soil	0.8615 ± 0.1718	0.9135 ± 0.0046	0.9936 ± 0.0024	0.9948 ± 0.0058	0.9855 ± 0.0076	0.9991 ± 0.0007	**0.9996 ± 0.0008**
Beet	0.9625 ± 0.0389	0.9848 ± 0.0016	0.9728 ± 0.0022	0.9712 ± 0.0110	**0.9929 ± 0.0028**	0.9849 ± 0.0078	0.9898 ± 0.0109
Wheat 2	0.8874 ± 0.0758	0.9853 ± 0.0028	**0.9942 ± 0.0015**	0.9577 ± 0.0259	0.9827 ± 0.0069	0.9404 ± 0.0166	0.9761 ± 0.0263
Wheat 3	0.8478 ± 0.1554	0.9956 ± 0.0014	0.9875 ± 0.0019	0.9799 ± 0.0048	0.9853 ± 0.0104	0.9926 ± 0.0032	**0.9970 ± 0.0034**
Stembeans	0.9423 ± 0.0593	0.9857 ± 0.0021	0.9902 ± 0.0022	0.9806 ± 0.0220	0.9964 ± 0.0029	0.9741 ± 0.0070	**0.9989 ± 0.0010**
Rapeseed	0.8558 ± 0.1639	0.9901 ± 0.0009	0.9028 ± 0.0026	0.9688 ± 0.0299	0.9452 ± 0.0396	0.9521 ± 0.0116	**0.9898 ± 0.0037**
Wheat	0.8832 ± 0.0952	**0.9994 ± 0.0002**	0.9823 ± 0.0014	0.9643 ± 0.0223	0.9847 ± 0.0177	0.9604 ± 0.0066	0.9889 ± 0.0067
Buildings	0.8588 ± 0.2273	0.9745 ± 0.0027	0.9640 ± 0.0023	0.9801 ± 0.0170	0.9886 ± 0.0091	0.9801 ± 0.0116	**0.9973 ± 0.0000**
Potatoes	0.8069 ± 0.2284	0.9910 ± 0.0007	0.9467 ± 0.0045	0.9538 ± 0.0371	**0.9971 ± 0.0015**	0.9772 ± 0.0066	0.9673 ± 0.0619
AA	0.8418 ± 0.0568	0.9590 ± 0.0023	0.9662 ± 0.0002	0.9776 ± 0.0042	0.9813 ± 0.0044	0.9716 ± 0.0047	**0.9910 ± 0.0040**
Kappa	0.8324 ± 0.0659	0.9749 ± 0.0027	0.9635 ± 0.0022	0.9731 ± 0.0056	0.9807 ± 0.0040	0.9674 ± 0.0066	**0.9882 ± 0.0057**
OA	0.8457 ± 0.0610	0.9769 ± 0.0023	0.9666 ± 0.0033	0.9753 ± 0.0052	0.9822 ± 0.0037	0.9700 ± 0.0060	**0.9892 ± 0.0052**

The bold values indicate the best performance in that category or evaluation indicator.

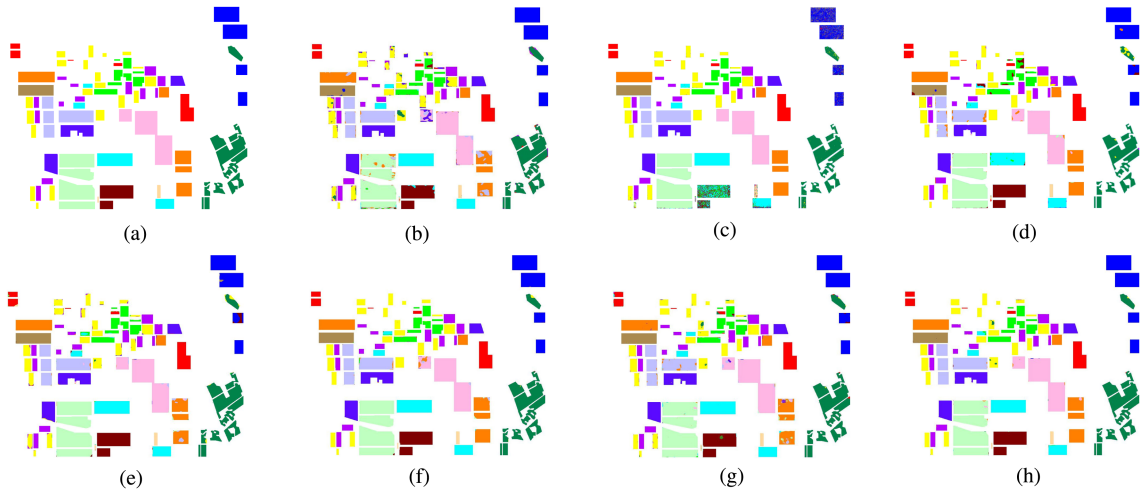


Fig. 8. (a) Ground truth map. (b) ResNet. (c) CV-FCN. (d) CV-3D-CNN. (e) SViT. (f) CCT. (g) MCPT. (h) Proposed method.

model in all aspects, which is instructive for the results of the experiments.

D. Results

In this section, the results and detailed analysis of the comparison experiments are displayed and described separately according to the datasets. The analysis of the experimental results for each dataset consists of an objective analysis of the evaluation metrics as well as a subjective judgment of the predicted images, which makes the results of the analysis more realistic and credible. Since the experiment uses a five-fold cross-validation method, the objective evaluation metrics show the mean and the corresponding standard deviation of the five-fold cross-validation. In the subjective judgments of the prediction result images, the corresponding colors are used directly to illustrate the predicted classification results to make the analysis more intuitive and clearer.

1) Analysis of the experimental results of the AIRSAR Flevoland dataset: The detailed results of the comparison experiments are given in Table III and the predicted classification results for all pixels are shown in Fig. 9. The overall results of the classical network model ResNet are poor compared to other methods, and each metric has a significant gap with other methods. Moreover, the standard deviation of three evaluation metrics of the method is large, which indicates the poor stability of the method. The CV-FCN and CV-3-D-CNN methods are based on the improvement of CNN for the characteristic that PolSAR data are complex-valued, so all their evaluation indicators are more advantageous compared with ResNet. The CV-FCN achieves the highest classification accuracy in the comparison experiments on the wheat category, and the CV-3-D-CNN shows the best

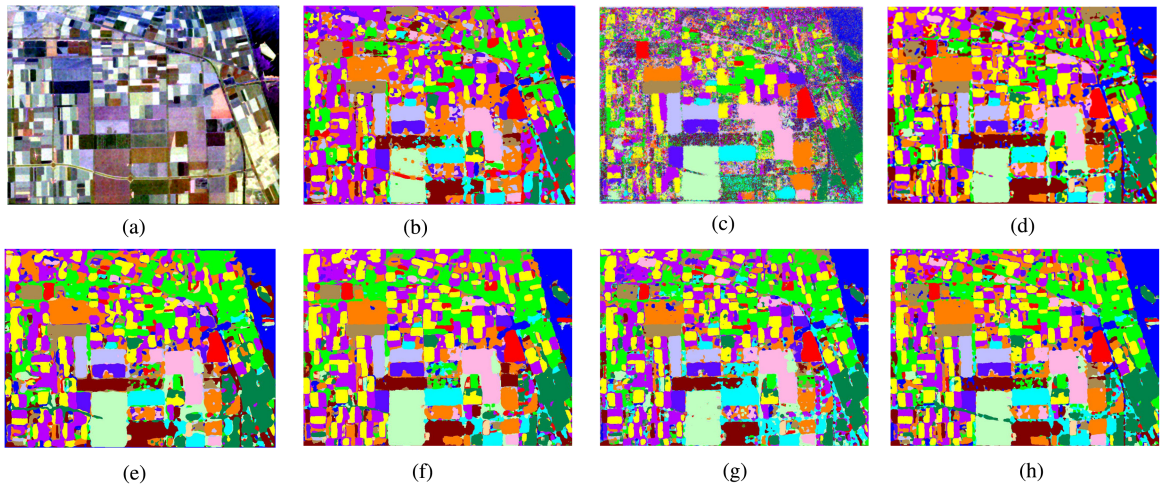


Fig. 9. (a) Pauli-RGB image. (b) ResNet. (c) CV-FCN. (d) CV-3D-CNN. (e) SViT. (f) CCT. (g) MCPT. (h) Proposed method.

classification accuracy in both categories, forest and wheat 2. SViT is the first ViT-based method introduced for PolSAR image classification. The method has a good performance with high indicators and balanced classification accuracy in each class, where the highest classification accuracy is achieved in the barley class. In the methods based on the combination of CNN and ViT, both CCT and MCPT have good classification performance and perform better than most of the single methods. It demonstrates to some extent that the combination of CNN and ViT has a positive effect on the classification performance. The method proposed further supports the feasibility of this idea. The proposed method reaches the highest OA on the AIRSAR Flevoland dataset. It not only obtains the best performance on all three evaluation metrics, but also accomplishes the highest classification accuracy on all ten categories. The building class has the least labeled data, which means that the proposed method still maintains a good performance in less sample classification. Besides, the standard deviations of the metrics of the proposed method are all small, which means that the model is well stabilized and has low sensitivity to the data.

Because the AIRSAR Flevoland dataset has less labeled data, additional images of the prediction results with only the labeled portion are shown here and compared to the ground truth map as shown in Fig. 8. It can better illustrate the predictive effectiveness of the model on the labeled data, and fully demonstrate the ability of the model to learn the features of the PolSAR data. As can be seen from Fig. 8(b), ResNet does not predict the labeled regions well, and there are many cases of prediction errors. Not only the edge portions, but also within the regions are similarly full of prediction errors. It is also consistent with the performance of the objective evaluation indicators of the method. Fig. 8(c) illustrates the prediction result of CV-FCN. It can be seen that most of the regions are able to be correctly predicted. However, there appear to be many spots of other colors in the cyan area at the bottom of the image and the blue area in the upper right corner. It reveals that the method is less effective in predicting individual classes and does not fully capture the characteristics of the data. The green part of the upper right corner in Fig. 8(d) shows a more severe prediction error problem, and other regions also have edge prediction errors that propagate to the interior. In Fig. 8(e), it is basically the edge predictions that are faulty. Fig. 8(f) presents the CCT works well. The predictions are largely correct except for a few isolated areas in the middle section. Most of the regions in Fig. 8(g) are predictable, and only a few areas show spotty conditions. Fig. 8(h) shows the prediction result of the proposed method. It is evident that there are basically very few cases of prediction errors, which is close to the ground truth map. The results indicate that the proposed method has the best performance for the prediction of the labeled portion and is able to learn the PolSAR image features well.

In addition to the prediction of labeled regions, PolSAR image classification places more importance on the prediction results for unlabeled parts. Fig. 9 exhibits the full pixel prediction results of each method for the AIRSAR Flevoland dataset and compares them with the Pauli-RGB image. From the classification prediction image Fig. 9(b), it can be seen that the boundary of each region in the prediction image of ResNet is not clear enough, and a lot of places are stuck together and cannot be clearly distinguished. The CV-FCN prediction image Fig. 9(c) illustrates the classification performance for the unlabeled region is inadequate as there are multiple colors mixed together, leading to difficulties in distinguishing the specific region category. It is clear from Fig. 9(d) that the CV-3-D-CNN method has a better classification effect. The unlabeled part can also be predicted well. However, the blending of colors in specific areas within the image indicates a range of categories. It is obvious from Fig. 9(e) that although SViT method has a high level of accuracy performance, the actual presentation of prediction image is not satisfactory. The edges of most of its areas appear diffuse, and the different categories intermingle with each other in the lack of distinct boundaries. The CCT prediction image presented in Fig. 9(f) suggests that most of the regions are well predicted with the correct category and the purity within the region is high. Yet, the predicted colors of the unlabeled location in the upper left corner of its image are fused together and the original region delineation is also lost. The prediction result map Fig. 9(g)

TABLE IV
OBJECTIVE EVALUATION INDICATORS OF SEVEN METHODS ON THE RADARSAT-2 SAN FRANCISCO DATASET

	ResNet	CV-FCN	CV-3-D-CNN	SViT	CCT	MCPT	Proposed
Water	0.7784 ± 0.4539	0.9921 ± 0.0009	0.9938 ± 0.0029	0.9985 ± 0.0019	0.9997 ± 0.0003	0.9766 ± 0.0471	**0.9969 ± 0.0060**
Vegetation	0.9363 ± 0.0321	0.9522 ± 0.0067	**0.9554 ± 0.0019**	0.9223 ± 0.0378	0.8991 ± 0.0065	0.9064 ± 0.0182	0.9211 ± 0.0571
High-density urban	0.8765 ± 0.1152	0.9311 ± 0.0056	0.9495 ± 0.0008	0.9629 ± 0.0102	0.9611 ± 0.0051	0.9583 ± 0.0089	**0.9665 ± 0.0123**
Developed	0.9422 ± 0.0381	0.9672 ± 0.0046	0.9345 ± 0.0017	0.9628 ± 0.0177	0.9626 ± 0.0024	0.9238 ± 0.0172	**0.9635 ± 0.0142**
Low-density urban	0.9416 ± 0.0647	0.8955 ± 0.0059	0.9224 ± 0.0017	0.9310 ± 0.0131	0.9449 ± 0.0065	0.9428 ± 0.0122	**0.9371 ± 0.0571**
AA	0.8950 ± 0.0879	0.9476 ± 0.0036	0.9511 ± 0.0008	0.9555 ± 0.0058	0.9535 ± 0.0015	0.9416 ± 0.0109	**0.9570 ± 0.0087**
Kappa	0.8180 ± 0.2362	0.9537 ± 0.0017	0.9532 ± 0.0015	0.9542 ± 0.0075	0.9541 ± 0.0017	0.9366 ± 0.0308	**0.9555 ± 0.0125**
OA	0.8536 ± 0.2078	0.9542 ± 0.0041	0.9629 ± 0.0023	0.9682 ± 0.0052	0.9681 ± 0.0012	0.9556 ± 0.0223	**0.9690 ± 0.0087**

The bold values indicate the best performance in that category or evaluation indicator.

of MCPT seems to be more neatly divided into regions as a whole, but the boundaries of each region are blurred and can be encroached by other colors. It is also less pristine within most regions and more forecasting errors occur. While the prediction result map Fig. 9(h) of the proposed method is able to distinguish each region well, the boundaries are also clearer, and the internal purity of most of the regions is also higher. It is evident from the prediction images that the CNN-based method is more accurate for localized prediction, i.e., each region has a higher degree of internal purity. But, these methods are not clear enough for the overall area delineation, which is manifested in the lack of clear boundaries. Whereas the ViT-based approach has a better grasp of the overall structure, it is slightly less specific to each part. The combined methods demonstrate the strengths of both approaches to various degrees. However, there is a lack of balance between CNN and ViT, which can lead to differences in emphasis. The CCT approach is more localized, while MCPT is more concerned with the global picture. The proposed method in this article better balances the advantages of the two. And local information is preserved as well as having a good overall view.

2) Analysis of the experimental results of the RADARSAT-2 San Francisco dataset: As can be seen in Table IV, there is still a large gap between the performance of ResNet on this dataset compared to other methods. And the standard deviation of its indicators is poor, which again points to the instability of the model. CV-FCN and CV-3-D-CNN perform moderately well on this dataset, although CV-3-D-CNN performs best in classifying the vegetation class. SViT and CCT have a dominant performance on this dataset and both have high OA. Yet the CCT model is more stable as can be seen from the standard deviation. MCPT performs poorly on this dataset, which is below average and has a large standard deviation. It suggests that this dataset is a challenge to its feature learning capability. The proposed method still retains its advantages. Not only does it achieve the highest classification accuracy in all four categories, but it also reaches the highest level among the compared methods in all three evaluation metrics. The results again illustrate that the proposed method possesses strong feature extraction capability. Regardless of the size of the PolSAR image, the model is able to learn the feature information contained in it very well.

Since there are fewer unlabeled areas in the RADARSAT-2 San Francisco dataset, the prediction results for the labeled portion are not shown separately. As seen from the prediction image Fig. 10(b), ResNet has more classification errors in the vegetation class, i.e., many other colors appear in green. Fig. 10(c) shows a figure of the prediction results of CV-FCN. The red color appears more at the left edge of the land, which is largely absent from the other prediction images. Moreover, it misses some of the green lines in the yellow area that are present in all the other images. The blue areas in the land appear in Fig. 10(d) that are not present in the other images. The result is more consistent with what is shown in the Pauli-RGB image. CV-3D-CNN shows a better prediction ability that other methods do not have, and with less clutter in the blue and yellow regions. However, it appears more yellow inside the red region, indicating some problems with the prediction of the high-density urban class. The top portion of yellow in Fig. 10(e) is encroached upon by blue and cannot be connected to the formation. It is because SViT only focuses on global information resulting in the loss of local information. The prediction image Fig. 10(f) of CCT compensates for the problem of localized information loss to some extent. But, it is too aggressive in its prediction of the low-density urban class, even though purple parts appear inside the red regions. For the MCPT prediction image Fig. 10(g), there is nothing wrong with the overall structure. Yet, each color area displays a lot of stray portions of other colors, especially the red portions small and numerous in the yellow areas. The prediction image Fig. 10(h) of the proposed method is purer within each color region. There are some other stray colors as well, but they are already less present compared to the other images. For the overall structure as well as local details can be better shown.

3) Analysis of the experimental results of the ESAR Oberpfaffenhofen dataset: Table V shows that ResNet has the worst feature learning ability for this dataset. In particular, the method has poor classification performance for the open area category and is very unstable. Both CV-FCN and CV-3-D-CNN perform better on this dataset. CV-3-D-CNN even achieves the highest classification accuracy in the wood land category. It indicates that the use of complex values does have a very positive effect on PolSAR image classification. Of the three ViT-based methods, only CCT achieves an overall average classification accuracy of more than 90%. The other two methods performed poorly and only moderately well. The presentation of the proposed methodology continues to be superior. The best classification results are obtained on both categories and the three evaluation metrics are also the highest among the compared methods.

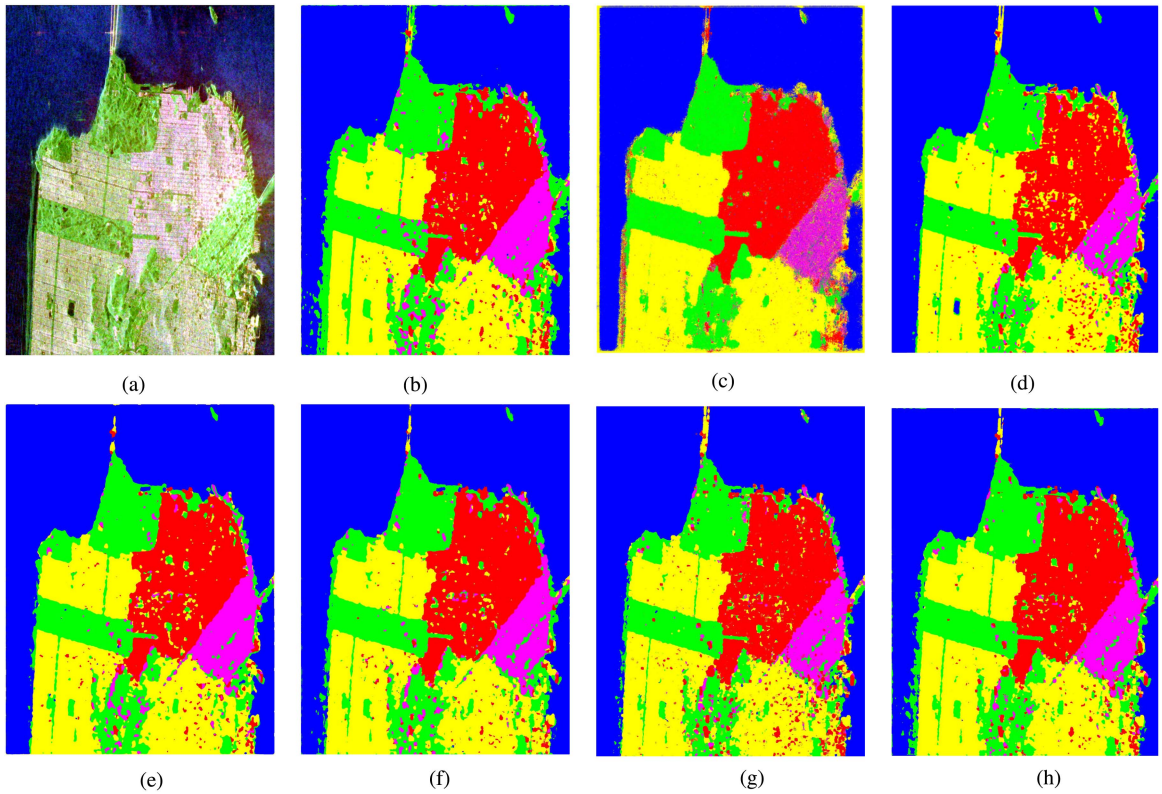


Fig. 10. (a) Pauli-RGB image. (b) ResNet. (c) CV-FCN. (d) CV-3D-CNN. (e) SViT. (f) CCT. (g) MCPT. (h) Proposed method.

TABLE V
OBJECTIVE EVALUATION INDICATORS OF SEVEN METHODS ON THE ESAR OBERPFAFFENHOFEN DATASET

	ResNet	CV-FCN	CV-3-D-CNN	SViT	CCT	MCPT	Proposed
Built-up area	0.9284 ± 0.0621	0.9722 ± 0.0004	0.9104 ± 0.0020	0.9289 ± 0.0459	0.9276 ± 0.0195	0.9206 ± 0.0294	**0.9316 ± 0.0561**
Wood land	0.8205 ± 0.2513	0.7810 ± 0.0016	**0.9333 ± 0.0010**	0.7524 ± 0.0732	0.7938 ± 0.0384	0.6679 ± 0.0649	0.7990 ± 0.0728
Open area	0.5205 ± 0.3509	0.9256 ± 0.0030	0.9263 ± 0.0015	0.9444 ± 0.0202	0.9438 ± 0.0215	0.9649 ± 0.0089	**0.9667 ± 0.0189**
AA	0.7565 ± 0.1051	0.8929 ± 0.0009	0.9233 ± 0.0006	0.8752 ± 0.0115	0.8884 ± 0.0117	0.8512 ± 0.0124	**0.8991 ± 0.0053**
Kappa	0.5306 ± 0.2116	0.8243 ± 0.0026	0.8601 ± 0.0110	0.8205 ± 0.0203	0.8365 ± 0.0112	0.7986 ± 0.0144	**0.8605 ± 0.0065**
OA	0.6742 ± 0.1756	0.9009 ± 0.0008	0.9142 ± 0.0020	0.8939 ± 0.0127	0.9035 ± 0.0068	0.8829 ± 0.0079	**0.9184 ± 0.0042**

The bold values indicate the best performance in that category or evaluation indicator.

Although the proposed method performs slightly worse in the wood land category, its overall performance remains high, which demonstrates the excellent performance of the proposed method.

The prediction image Fig. 11(b) of ResNet can clearly see a lot of green parts in the red region, which confirms the poor classification performance of the method on the Wood Land class. But it still has some value with fewer stray colors within the green and yellow areas. Fig. 11(c) illustrates that CV-FCN is accurate in predicting the red region, and only this method predicts the red region in the lower half of the image more completely. Overall, each area seems more complete, and there are fewer large areas of other colors in the interior. CV-3D-CNN predicts each region more accurately, and the green part in the red region in Fig. 11(d) is small and few. The prediction image Fig. 11(e) of SViT is not much different from Fig. 11(d) and both appear red at the edges of the green part. Whereas the CCT improved this phenomenon in its predicted image, as shown in Fig. 11(f). However, its yellow area appears with red blotches. The yellow area in Fig. 11(g) of MCPT displays the same red spots as Fig. 11(f) and it shows more green parts in the red area with more stray colors inside the area. The prediction image Fig. 11(h) of the proposed method is purer in each region and less of other color patches appear. Besides, the overall structure is clear and presents a good categorization of both labeled and unlabeled sections.

In summary, the results of the experiments on the three datasets are mutually corroborated in two ways through objective analysis and subjective judgments of the experimental results. It is finally evidenced that the proposed method in this article has higher classification accuracy and better performance on all three datasets. By comparing with ResNet, CV-FCN, CV-3-D-CNN, and SViT which are based on a single method, the proposed method based on the combination of the two is effective and superior. Therefore, it can better extract the rich

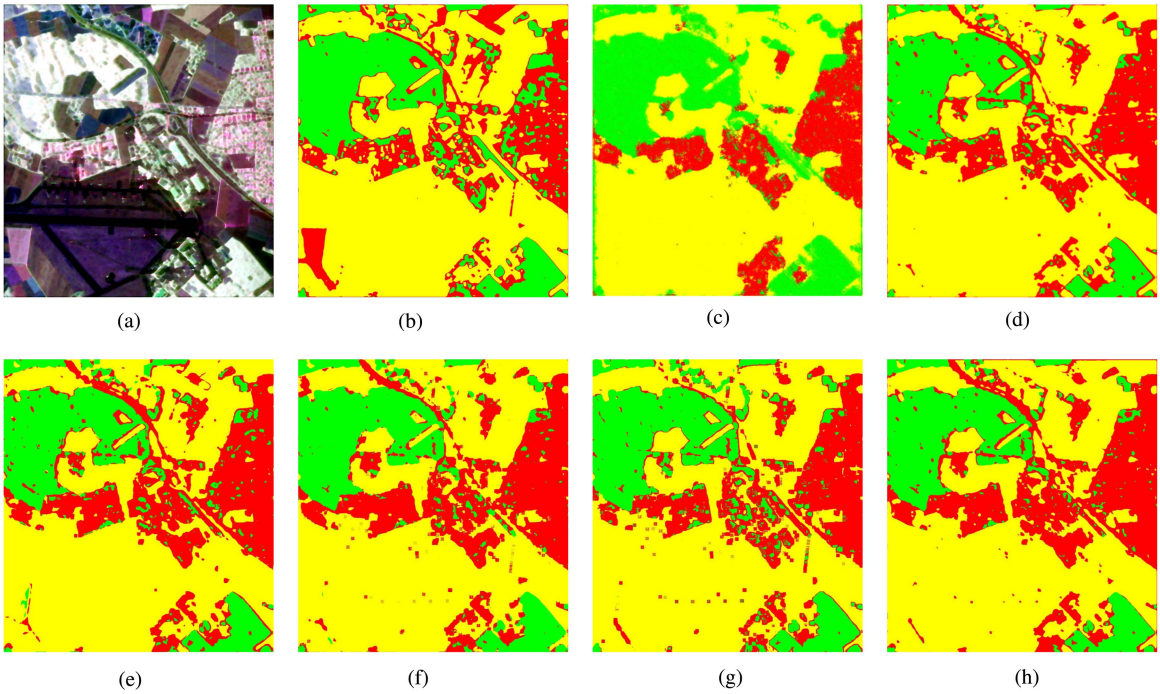


Fig. 11. (a) Pauli-RGB image. (b) ResNet. (c) CV-FCN. (d) CV-3-D-CNN. (e) SViT. (f) CCT. (g) MCPT. (h) Proposed method.

TABLE VI
EXPERIMENTAL RESULTS OF ABLATION STUDY ON THREE DATASETS

Experiment	OA		
	Flevoland	San Francisco	Oberpfaffenhofen
(1)ViT	97.80	96.37	91.46
(2)ViT+ET	99.04	97.12	92.11
(3)ViT+DGA_CAFF	98.10	96.76	91.58
(4)ViT+ET+DGA_CAFF	**99.36**	**97.63**	**93.04**

The bold values indicate the best performance in that category or evaluation indicator.

information contained in PolSAR data, and has a stronger feature representation learning capability. Further compared to CCT and MCPT, which are methods based on the combination of CNN and ViT, the proposed method better balances the characteristics so that it can better utilize the advantages of them.

IV. DISCUSSION

In this section, several factors that affect the performance of the model are discussed with the following three main components, namely ablation experiments, amount of training data, and granularity selection. The impact of these factors on model performance is explored through specific experimental designs and results, which will provide a more comprehensive and rigorous assessment of the proposed method.

A. Ablation Study

To better validate the effectiveness of the proposed method, ablation experiments are performed on the three datasets. It is worth noting that MGA and CAFF have to be used together. Therefore, when conducting ablation experiments, MGA_CAFF indicates both modules are activated at the same time. Meanwhile, in order to make the experimental results more intuitive, only OA of the three evaluation indicators is used in the ablation experiments for the evaluation of the classification performance. Table VI shows the results of the ablation experiments and each experiment is numbered in the table in order to facilitate the analysis of the results.

Experiment (1) in Table VI indicates that only the original ViT is used for the experiment, and it can be seen that the original ViT already has a relatively good classification accuracy. Experiment (2) adds the ET module to ViT. It is observed that the addition of the ET module is more helpful in improving the classification performance of the model. The OAs on the three datasets are improved by 1.24%, 0.75%, and 0.65%, respectively. The results of this experiment explain that the ET module can well extract the local features of the data to improve the final classification performance of ViT. Experiment (3) is the introduction of the proposed DGA_CAFF module based on ViT. The introduction

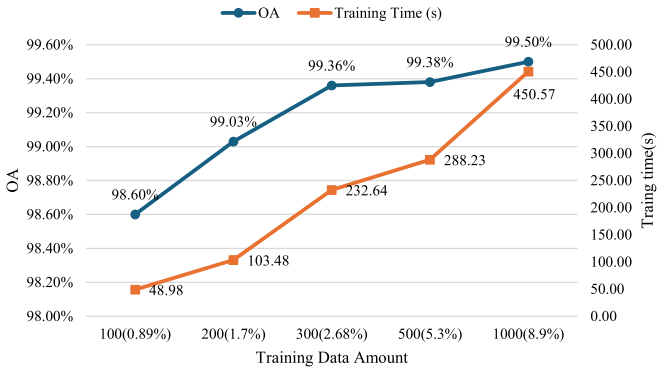


Fig. 12. Experimental plot about the impact of training data amounts.

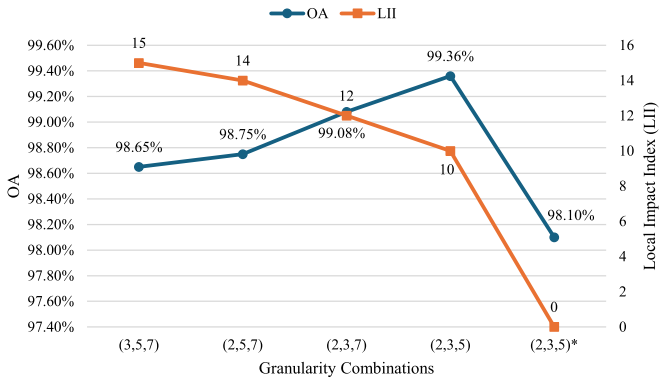


Fig. 13. Experimental plot about the influence of granularities selection and local features.

of this module can also provide an improvement in the model classification accuracy. Although the boost over the ET module is not as great, there is a degree of advantage over the ViT. Experiment (4) is the proposed method. A significant improvement in the overall classification performance of the model can be seen when the proposed three modules are used together. Compared to the original ViT, the OAs of the three datasets are improved by 1.56%, 1.26%, and 1.58%, respectively.

To summarize, each of the modules proposed in this article provides some performance improvement to the original ViT. In addition, the proposed method better combines the advantages of CNN and ViT, which makes the classification performance on all three datasets more advantageous. The results of the ablation experiments further demonstrate the feasibility as well as the superiority of the proposed method to better accomplish the PolSAR image classification task.

B. Impact of Training Data Amount

It is well known that the training effectiveness of deep learning models is limited by the amount of training data [59]. For the original ViT, a large amount of labeled data is required if it is to be adequately trained [60]. But labeling the PolSAR dataset pixel by pixel depends heavily on expert knowledge and is very time consuming. In most cases, labeling information can be acquired for only a few pixels in the PolSAR image. Therefore, an experiment is designed to train the model using different amounts of data and the model is evaluated on two aspects, OA and training time. The effect of the amount of training data on the proposed method is explored through this experiment.

In specific experiments, the model is trained on the AIRSAR Flevoland dataset using five different data volumes, 100 per category (0.89%), 200 per category (1.7%), 300 per category (2.68%), 500 per category (5.3%), and 1000 per category (8.9%), respectively. The parentheses are the proportion of training data volume to all labeled data, and the experimental results for different training data volumes are shown in Fig. 12. As it is seen, although the OA of the model tends to increase as the amount of training data increases, at the same time the training time of the model is also rising. When the amount of training data goes from 100 to 300 per class, the OA improves significantly, although simultaneously the training time also grows dramatically. Such results are intuitive. However, when the amount of training data is added from 300 to 1000 per class, the OA curve does not improve much. Instead, the training time increases sharply. It indicates that increasing the amount of data has a large effect on the OA of the model when the amount of training data is less than 300 per class, but this effect becomes small when it is more than 300 per class. As a result, the proposed model requires only 300 labeled data per class for training on the AIRSAR Flevoland dataset, which is more practical for the PolSAR image classification task with less labeled data.

C. Influence of Granularities Selection and Local Features

This article presents a multigranularity approach that incorporates local features. Therefore, the choice of combinations for different granularities is very much in need of discussion. Not all granularity combinations lead to better feature information. Several different granularity combinations are validated on the AIRSAR Flevoland dataset to investigate the effect of different granularity combinations and local features on the performance of the proposed model.

The granularity combinations used for the experiments are (3, 5, 7), (2, 5, 7), (2, 3, 7), and (2, 3, 5). Again, OA is used as an evaluation metric for the model. Meanwhile, according to the design of the model, the granularity size is also closely related to the depth of the feature map output by the ET module. The greater the granularity, the deeper the feature map is needed and the more abstract the features included in it. To better analyze the impact of the incorporation of local features on the final classification performance of CNN combined with ViT, the concept of LII is introduced into the experiments. The value of LII is the sum of the three granularity sizes. Due to the nature of the proposed ET module, when the value of LII is larger, it indicates that the ETs used come from a deeper extraction of features from the module and contain more local information. This affects the global nature of the subsequent ViT and leads to an influence on the model performance. In particular, the LII value is 0 when the ET module is not used. From the results of the first four experiments in Fig. 13, it is clear that as LII decreases, OA is gradually increasing, which is negatively correlated with each other. However, it does not mean that the inclusion of

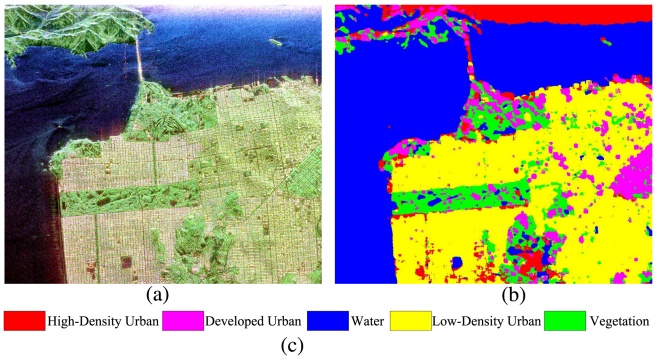


Fig. 14. Model generalization performance study result. (a) Pauli-RGB image of AIRSAR San Francisco dataset. (b) Prediction result image of AIRSAR San Francisco dataset using the model trained on RADARSAT-2 San Francisco dataset. (c) Legend of RADARSAT-2 San Francisco dataset.

localized features results in lower OA. (2, 3, 5)* denotes the result of the experiment using the granularity (2, 3, 5) but with the ET module removed, i.e., the result of the ablation Experiment (3). It has an LII value of 0, however, the OA gets lower and is lower than the results of all the experiments containing localized features.

The experimental results prove that the incorporation of local features is beneficial to enhancing the feature representation and classification performance of ViT. On the contrary, adding too many localized features will also degrade the classification performance of ViT. A balance between local and global features is needed. Of course, this choice is not fixed, it also depends on the input image size. When the input image is larger, the size of the output feature map of the convolution layer will have more choices. It means that there will be more combinations of granularity, but it will also increase the computational overhead of the model. When the input image is much smaller, no more convolutional computations can be performed and the size of the feature map output from the convolutional layer is limited. As a result, the combination of granularities that can be selected will also be limited, but this will relatively reduce the computational cost of the model. Based on the above-mentioned considerations and experimental validation, the granularity combination of this balance point in this article is (2, 3, 5).

D. Model Generalization Performance Study

Although three different datasets have been used to validate the performance of the proposed method in previous comparison experiments, an additional AIRSAR San Francisco dataset is used for the generalization performance study to further illustrate the generalization ability of the model. The Pauli-RGB image of the dataset is shown in Fig. 14(a). It was acquired by the AIRSAR platform in 1989 with a spatial resolution of 10 m and an original image size of 900×1024. The dataset contains five land cover types including mountain, water, urban, vegetation, and bare soil.

In this experiment, a model trained on the RADARSAT-2 San Francisco dataset is utilized to make predictions directly on the AIRSAR San Francisco dataset. The predicted result is presented in Fig. 14(b), where the different colors represent the land cover types given in Fig. 14(c). For the classes water and vegetation, which are common to both datasets, the model is able to make good predictions. The blue portion of Fig. 14(b) is largely predicted correctly and the green portion is mostly predicted. However, it is difficult to correctly predict land cover types that have not been learned by the model, such as mountain and bare soil. But as can be seen by the Pauli-RGB image comparison, most of the bare soil is predicted to be yellow, i.e., the low-density urban category. It suggests a large similarity between the two land cover types. For categories that the model has not learned, it predicts the most similar land cover types, which is understandable and intuitive. The prediction results for the Urban category reveal that although the AIRSAR San Francisco dataset does not further differentiate this category, the model still breaks it down according to existing knowledge.

The experimental result indicates that the proposed method has good generalization performance. For different datasets and land cover types, the model is first able to predict well for the same categories. Second, the model for completely different feature classes is also able to predict them as the class with the most similar characteristics. Finally, for categories with more meaning, the model is capable of subdividing the category area based on existing knowledge as well. It further illustrates the superior performance of the proposed method with a strong feature learning representation and strong generalization.

TABLE VII
FLOPS AND PARAMS FOR SEVEN METHODS

	FLOPs	Params
ResNet	61.523M	853.427K
CV-FCN	2.273K	966.165K
CV-3D-CNN	126.093M	1.869M
SViT	45.503M	101.379K
CCT	14.880M	257.341K
MCPT	74.914M	4.101M
Proposed	237.832M	6.178M

E. Research on Model Computation Cost

Due to the fact that both the proposed method and the comparison methods used are based on deep learning, it is necessary to discuss the computational costs of them. To measure the computational effort of these methods, the number of floating point operations (FLOPs) and the number of parameters (Params) are chosen as metrics. FLOPs represents the computational complexity of the model, while Params refers to the number of model parameters.

The FLOPs and Params for each method are given in Table VII. As can be seen from Table VII, the computation and number of parameters of the proposed model are the largest. It is related to the structural complexity of the proposed model. The aim of this article is to solve the problem that a single network structure cannot fully extract the rich feature information contained in PolSAR data, and that a large amount of labeled data

is required for the training of existing deep learning methods. So, the proposed method employs a network structure that uses a fusion of CNN and ViT. Inevitably, it makes the model more complex and increases the computational and parametric quantities of the model, but it can solve the above-mentioned problems well. In comparison with other models with less computational and parametric quantities, the proposed model is better able to extract more comprehensive information about the PolSAR data and requires only a small amount of labeled data to complete the training.

V. Conclusion

In order to solve the problem that the existing single network makes it difficult to extract the complex and rich information contained in PolSAR data, and that deep learning methods require a large amount of PolSAR labeled data, this article explores more deeply how CNN and ViT can be effectively fused to extract PolSAR image features. Therefore, a multigranularity hybrid CNN-ViT model based on ETs and cross-attention is proposed for PolSAR image classification. The ET module can effectively extract the local information of the PolSAR image and add them as ETs to the original patch tokens in the ViT branch, which is useful to make up for the lack of local information neglected by ViT. The application of MGA further improves the richness and confidence of the extracted features of the model. In addition, through the CAFF module, the features of the two branches can be better complementarily fused to more effectively represent the feature information contained in the PolSAR image. Experimental results demonstrate the effectiveness of the proposed method. Not only can the rich feature information contained in PolSAR data be well extracted, but also the labeled data required for training is very little, and the complementary fusion between CNN and ViT is well accomplished to achieve the classification of PolSAR images. However, the proposed method still has some shortcomings. There is only a consideration of the feature extraction process of the PolSAR data, and a more in-depth study of the data distribution is lacking. In addition, the proposed model is complex and lacks consideration of computational costs. In subsequent work, the research focus will be on the study of PolSAR data distribution and the optimization of the computational cost of the model to better explore the implementation of PolSAR image classification.

References

[1] Y. K. Chan and V. C. Koo, "AN introduction to synthetic aperture radar (SAR)," *Prog. Electromagnetics Res. B*, vol. 2, pp. 27–60, 2008.

[2] R. Bamler, "Principles of synthetic aperture radar," *Surv. Geophys.*, vol. 21, no. 2, pp. 147–157, Mar. 2000.

[3] F. Ma, X. Sun, F. Zhang, Y. Zhou, and H.-C. Li, "What catch your attention in SAR images: Saliency detection based on soft-superpixel lacunarity cue," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

[4] T. Strozzi, U. Wegmuller, G. Bitelli, and V. Spreckels, "Land subsidence monitoring with differential SAR interferometry," *Photogrammetric Eng. Remote Sens.*, vol. 67, no. 11, pp. 1261–1270, Nov. 2001.

[5] W. Stecz and K. Gromada, "UAV mission planning with SAR application," *Sensors*, vol. 20, no. 4, Jan. 2020, Art. no. 1080.

[6] Z. Chen, Y. Zhang, B. Guindon, T. Esch, A. Roth, and J. Shang, "Urban land use mapping using high resolution SAR data based on density analysis and contextual information," *Can. J. Remote Sens.*, vol. 38, no. 6, pp. 738–749, Jan. 2013.

[7] G. Milani et al., "Robust quantification of riverine land cover dynamics by high-resolution remote sensing," *Remote Sens. Environ.*, vol. 217, pp. 491–505, Nov. 2018.

[8] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[9] H. Zebker and J. Van Zyl, "Imaging radar polarimetry: A review," *Proc. IEEE*, vol. 79, no. 11, pp. 1583–1606, Nov. 1991.

[10] S.-W. Chen, "Polarimetric coherence pattern: A visualization and characterization tool for PolSAR data investigation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 286–297, Jan. 2018.

[11] H. Parikh, S. Patel, and V. Patel, "Classification of SAR and PolSAR images using deep learning: A review," *Int. J. Image Data Fusion*, vol. 11, no. 1, pp. 1–32, Jan. 2020.

[12] W. Cameron and L. Leung, "Feature motivated polarization scattering matrix decomposition," in *Proc. IEEE Int. Conf. Radar*, 1990, pp. 549–557.

[13] S. Cloude, "Target decomposition theorems in radar scattering," *Electron. Lett.*, vol. 21, no. 1, pp. 22–24, Jan. 1985.

[14] S. Cloude and E. Pottier, "An entropy based classification scheme for land applications of polarimetric SAR," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 1, pp. 68–78, Jan. 1997.

[15] E. Pottier, "Dr. J. R. Huynen's main contributions in the development of polarimetric radar techniques and how the 'Radar targets phenomenological concept' becomes a theory," in *Proc. SPIE*, SPIE, 1993, pp. 72–85.

[16] A. P. Doulgeris, S. N. Anfinsen, and T. Eltoft, "Automated non-Gaussian clustering of polarimetric synthetic aperture radar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3665–3676, Oct. 2011.

[17] L. I. Lan, C. Erxue, L. I. Zengyuan, F. Qi, and Z. Lei, "K-wishart classifier for PolSAR data and its performance evaluation," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 41, no. 11, pp. 1498–1504, Nov. 2016.

[18] F. Zhang, X. Sun, F. Ma, and Q. Yin, "Superpixelwise likelihood ratio test statistic for PolSAR data and its application to built-up area extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 233–248, Mar. 2024.

[19] W. Hua, L. Liu, N. Sun, and X. Jin, "A CA-Based weighted clustering adversarial network for unsupervised domain adaptation PolSAR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[20] S. Fukuda and H. Hirosawa, "Support vector machine classification of land cover: Application to polarimetric SAR data," in *Proc. Scanning Present Resolving Future. Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2001, pp. 187–189.

[21] C. Li, X. Tian, Z. Li, E. Chen, and W. Zhang, "Retrieval of forest above ground biomass using automatic KNN model," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 4418–4421.

[22] L. Zhang, L. Sun, and W. M. Moon, "Feature extraction and classification of PolSAR images based on sparse representation," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 2798–2801.

[23] J. Wang et al., "Parameter selection of touzi decomposition and a distribution improved autoencoder for PolSAR image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 186, pp. 246–266, Apr. 2022.

[24] W. Xie et al., "POLSAR image classification via Wishart-AE model or Wishart-CAE model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3604–3615, Aug. 2017.

[25] A. Jamali, M. Mahdianpari, F. Mohammadimanesh, A. Bhattacharya, and S. Homayouni, "PolSAR image classification based on deep convolutional neural networks using wavelet transformation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[26] Y. Li, Y. Chen, G. Liu, and L. Jiao, "A novel deep fully convolutional network for PolSAR image classification," *Remote Sens.*, vol. 10, no. 12, Dec. 2018, Art. no. 1984.

[27] C. Yang, B. Hou, B. Ren, Y. Hu, and L. Jiao, "CNN-Based polarimetric decomposition feature selection for PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8796–8812, Nov. 2019.

[28] L. Wang, X. Xu, H. Dong, R. Gui, and F. Pu, "Multi-pixel simultaneous classification of PolSAR image using convolutional neural networks," *Sensors*, vol. 18, no. 3, Mar. 2018, Art. no. 769.

[29] L. Li, L. Ma, L. Jiao, F. Liu, Q. Sun, and J. Zhao, "Complex Contourlet-CNN for polarimetric SAR image classification," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107110.

[30] R. Shang, J. Wang, L. Jiao, X. Yang, and Y. Li, "Spatial feature-based convolutional neural network for PolSAR image classification," *Appl. Soft Comput.*, vol. 123, Jul. 2022, Art. no. 108922.

[31] H. Parikh, S. Patel, and V. Patel, "Modeling PolSAR classification using convolutional neural network with homogeneity based kernel selection," *Model. Earth Syst. Environ.*, vol. 9, pp. 3801–3813, Feb. 2023.

[32] M. Li, Q. Shen, Y. Xiao, X. Liu, and Q. Chen, "PolSAR image building extraction with G0 statistical texture using convolutional neural network and superpixel," *Remote Sens.*, vol. 15, no. 5, Jan. 2023, Art. no. 1451.

[33] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–22.
[34] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 9355–9366.
[35] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," 2021, *arXiv:2112.13492*.
[36] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 22–31.
[37] X. Liu, Y. Wu, W. Liang, Y. Cao, and M. Li, "High resolution SAR image classification using global-local network structure based on vision transformer and CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[38] X. Liu, Y. Wu, X. Hu, Z. Li, and M. Li, "A novel lightweight attention-discarding transformer for high-resolution SAR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[39] H. Dong, L. Zhang, and B. Zou, "Exploring vision transformers for polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[40] Q. Yin, Z. Lin, W. Hu, C. López-Martínez, J. Ni, and F. Zhang, "Crop classification of multitemporal PolSAR based on 3-D attention module with ViT," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[41] F. Li, C. Zhang, X. Zhang, and Y. Li, "MF-DCMANet: A multi-feature dual-stage cross manifold attention network for PolSAR target recognition," *Remote Sens.*, vol. 15, no. 9, Jan. 2023, Art. no. 2292.
[42] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017.
[43] M. Q. Alkhatib, M. Al-Saad, N. Aburaed, M. S. Zitouni, and H. Al-Ahmad, "PolSAR image classification using attention based shallow to deep convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 8034–8037.
[44] H. Jin, T. He, J. Shi, and S. Ji, "Combine superpixel-wise GCN and pixel-wise CNN for polsar image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 8014–8017.
[45] J. Shi, T. He, S. Ji, M. Nie, and H. Jin, "CNN-Improved superpixel-to-pixel fuzzy graph convolution network for PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.
[46] M. Zhao, Y. Cheng, X. Qin, W. Yu, and P. Wang, "Semi-supervised classification of PolSAR images based on co-training of CNN and SVM with limited labeled samples," *Sensors*, vol. 23, no. 4, Jan. 2023, Art. no. 2109.
[47] A. A. Aleissaee et al., "Transformers in remote sensing: A survey," *Remote Sens.*, vol. 15, no. 7, Jan. 2023, Art. no. 1860.
[48] Y. Cao, Y. Wu, M. Li, M. Zheng, P. Zhang, and J. Wang, "Multifrequency PolSAR image fusion classification based on semantic interactive information and topological structure," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
[49] Y. Dong and R. Hänsch, "Multimodal self-supervised learning for semantic analysis of PolSAR imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 1704–1707.
[50] F. Liu, "PolSAR image classification and change detection based on deep learning," Ph.D. dissertation, Xidian Univ., Xi'an, China, 2017.
[51] X. Liu, L. Jiao, F. Liu, D. Zhang, and X. Tang, "PolSF: PolSAR image datasets on San Francisco," in *Proc. Int. Conf. Intell. Sci.*, Cham, 2022, pp. 214–219.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[53] Y. Cao, Y. Wu, P. Zhang, W. Liang, and M. Li, "Pixel-wise PolSAR image classification via a novel complex-valued deep fully convolutional network," *Remote Sens.*, vol. 11, no. 22, Jan. 2019, Art. no. 2653.
[54] X. Tan, M. Li, P. Zhang, Y. Wu, and W. Song, "Complex-valued 3-D convolutional neural network for PolSAR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1022–1026, Jun. 2020.
[55] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.
[56] W. Wang, J. Wang, B. Lu, B. Liu, Y. Zhang, and C. Wang, "MCPT: Mixed convolutional parallel transformer for polarimetric SAR image classification," *Remote Sens.*, vol. 15, no. 11, Jan. 2023, Art. no. 2936.
[57] A. J. Alberg, J. W. Park, B. W. Hager, M. V. Brock, and M. Diener-West, "The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests," *J. Gen. Intern. Med.*, vol. 19, no. 5, pp. 460–465, May 2004.
[58] H. C. Kraemer, "Kappa Coefficient," in *Wiley StatsRef: Statistics Reference Online*. Hoboken, NJ, USA: Wiley, 2015, pp. 1–4.
[59] A. Mathew, P. Amudha, and S. Sivakumari, "Deep Learning Techniques: An Overview," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.*, Singapore, Springer, 2021, pp. 599–608.
[60] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your ViT? data, augmentation, and regularization in vision transformers," *Trans. Mach. Learn. Res.*, Apr. 2022, pp. 1–16.

Wenke Wang received the B.S. degree in computer science and technology from Henan Polytechnic University, Jiaozuo, China, in 2020. He is currently working toward the M.S. degree in computer technology with the School of Computer Science and Technology, Henan Polytechnic University.

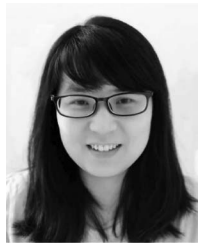
His research interests include image processing, feature extraction, deep learning, and polarimetric synthetic aperture radar image classification.

Jianlong Wang received the B.S. degree in computer science and technology from Zhengzhou University, Zhengzhou, China, in 2012, the M.S. degree in computer science and technology from Henan Polytechnic University, Jiaozuo, China, in 2015, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2021.

He is currently a Lecturer with the School of Computer Science and Technology, Henan Polytechnic University. His research interests include image processing, sparse representation, machine learning, deep learning, and polarimetric synthetic aperture radar image classification.

Dou Quan (Member, IEEE) received the B.S. degree in intelligence science and technology and Ph.D. degree in circuits and systems from Xidian University, Xian, China, in 2015 and 2021, respectively.

She is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. From 2019 to 2020, she was a Joint Ph.D. along with Prof. Jocelyn Chanussot with the Research Center of Inria Grenoble-Rhone-Alpes, Montbonnot-Saint-Martin, France. Her research interests include machine learning, deep learning and metric learning, image matching, image registration, change detection, and remote sensing image processing and interpretation.



Meijuan Yang (Member, IEEE) received the B.S. degree in electronic information science and technology from Xidian University, Xi'an, China, in 2010, the M.S. degree in signal and information processing from the University of Chinese Academy of Sciences, Xi'an, China, in 2013, and the Ph.D. degree in circuits and systems from Xidian University, in 2021.

She is currently an Associate Professor with the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an, China. Her research interests include deep learning and image processing.

Junding Sun received the B.S. degree in computer application and the M.S. degree in control theory and control engineering from Henan Polytechnic University, Jiaozuo, China, in 1998 and 2001, respectively. He received the Ph.D. degree in computer application from Xidian University, Xian, China, in 2005.

He is currently a Professor with the School of Computer Science and Technology, Henan Polytechnic University. His major interests include image processing, image retrieval and pattern recognition.

Bibo Lu received the B.S. and M.S. degrees in computational math and the Ph.D. degree in applied math from Jilin University, Changchun, China, in 2002, 2005, and 2008, respectively.

He is currently a Professor with the School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China. His research interests include image processing, video processing, and deep learning.