PolSAR-MPIformer: A Vision Transformer Based on Mixed Patch Interaction for Dual-Frequency PolSAR Image Adaptive Fusion Classification

Xinyue Xin, Ming Li, *Member, IEEE*, Yan Wu, *Member, IEEE*, Xiang Li, Peng Zhang, *Member, IEEE*, and Dazhi Xu

***Abstract*—Vision transformer (ViT) provides new ideas for polarization synthetic aperture radar (PolSAR) image classification due to its advantages in learning global-spatial information. However, the lack of local-spatial information within samples and correlation information among samples, as well as the complexity of network structure, limit the application of ViT in practice. In addition, dual-frequency PolSAR data provide rich information, but there are fewer related studies compared to single-frequency classification algorithms. In this article, we adopt ViT as the basic framework, and propose a novel model based on mixed patch interaction for dual-frequency PolSAR image adaptive fusion classification (PolSAR-MPIformer). First, a mixed patch interaction (MPI) module is designed for the feature extraction, which replaces the high-complexity self-attention in ViT with patch interaction intra- and intersample. Besides the global-spatial information learning within samples by ViT, the MPI module adds the learning of local-spatial information within samples and correlation information among samples, thereby obtaining more discriminative features through a low-complexity network. Subsequently, a dual-frequency adaptive fusion (DAF) module is constructed as the classifier of PolSAR-MPIformer. On the one hand, the attention mechanism is utilized in DAF to reduce the impact of speckle noise while preserving details. On the other hand, the DAF evaluates the classification confidence of each band and assigns different weights accordingly, which achieves reasonable utilization of the complementarity between dual-frequency data and improves classification accuracy. Experiments on four real dual-frequency PolSAR datasets substantiate the superiority of the proposed PolSAR-MPIformer over other state-of-the-art algorithms.**

***Index Terms*—Dual-frequency adaptive fusion, mixed patch interaction, PolSAR image classification, vision transformer.**

Manuscript received 2 February 2024; revised 19 March 2024; accepted 3 April 2024. Date of publication 10 April 2024; date of current version 22 April 2024. This work was supported in part by the Natural Science Foundation of China under Grant 62172321 and in part by the Civil Space Thirteen Five Years Pre-Research Project under Grant D040114. *(Corresponding authors: Ming Li; Yan Wu.)*

Xinyue Xin, Ming Li, Peng Zhang, and Dazhi Xu are with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China, and also with the Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China (e-mail: liming@xidian.edu.cn).

Yan Wu is with the Remote Sensing Image Processing and Fusion Group, School of Electronics Engineering, Xidian University, Xi'an 710071, China (e-mail: ywu@mail.xidian.edu.cn).

Xiang Li is with the Beijing Institute of Radio Measurement, Beijing 100854, China.

Digital Object Identifier 10.1109/JSTARS.2024.3386854

I. Introduction

POLARIMETRIC synthetic aperture radar (PolSAR) is a microwave remote sensing (RS) technology, which obtains information by transmitting and receiving electromagnetic waves under various polarization modes [1], [2], [3]. With the imaging capability in all-time and all-weather conditions, PolSAR has attracted a lot of attention in the field of RS recently. PolSAR image classification is an important research direction in the application of PolSAR images [4], [5], [6], [7], [8]. It assigns different category labels to pixels based on the PolSAR backscattering information, and plays an important role in various fields [9].

To achieve accurate terrain classification, researchers have made many attempts. Among them, traditional classification algorithms can be roughly categorized into two groups: approaches based on statistical models [10], [11], [12] and approaches based on scattering mechanisms [13], [14], [15]. Recently, deep learning (DL) algorithms represented by convolutional neural network (CNN) have made significant progress in PolSAR image classification tasks. For example, Zhou et al. [16] successfully applied CNN to PolSAR image classification task by using a 6-D real vector as input and learning advanced feature representations through two cascaded convolutional layers. The result verified the superiority of CNN in PolSAR image classification. In order to fully utilize the characteristics of PolSAR data, some studies have proposed to combine expert knowledge and target scattering mechanisms to design a polarimetric-feature-driven CNN model [17] and complex-valued neural network [18], [19]. However, due to the limited receptive field of convolution kernel, information interaction can only be carried out in short-range space, which limits the feature extraction ability of CNN.

Vision transformer (ViT) [20] is an emerging network architecture that can effectively process global-spatial information within samples through self-attention mechanisms. Therefore, ViT can be used to simulate long-range spatial relationships in PolSAR image samples. In order to apply ViT to synthetic aperture radar (SAR) and PolSAR image classification tasks, Dong et al. [21] proposed using multiheaded self-attention blocks instead of convolutional blocks for feature exaction, which made it possible for long-range information interaction. The experimental results have demonstrated the robustness of ViT in PolSAR

image classification tasks. On this basis, Liu et al. [22] proposed a global–local network structure (GLNS) for high-resolution SAR image classification, which learns local and global features through lightweight CNN and compact ViT, respectively, and fuses these two types of features through a fusion net. Fan et al. [23] designed a network based on swin transformer [24], and realized joint feature learning of local-spatial information and global-spatial information through shift window and local window self-attention. Furthermore, Wang et al. [25] proposed a semisupervised PolSAR image classification method based on ViT. It designs an autoencoder for pretrain process, and then finetunes the network with labeled samples.

Although ViT has powerful long-range feature processing capabilities and has been successfully applied in PolSAR image classification tasks, it still has some limitations. On the one hand, ViT only focuses on the global-spatial information within samples, neglecting the local-spatial information within samples and correlation information among samples. In PolSAR images, the pixel correlation in the local space within samples is strong, and pixels from the same category may be distributed in different samples. Therefore, the study of local-spatial information within samples and correlation information among samples is necessary for PolSAR image classification. On the other hand, the self-attention network in ViT structure has high complexity, which increases the difficulty of network training. For small-scale samples in PolSAR image classification, a highly complex self-attention network is not very necessary [26], [27], [28], [29].

The above algorithms are presented for single-frequency PolSAR image classification. With the development of multisensor technology, dual-frequency PolSAR data are gradually being applied for PolSAR image classification. Due to the fact that electromagnetic waves in different frequencies have different penetrability, PolSAR data in different frequency bands can obtain different feature representations for the same ground object [30], [31], [32]. For example, PolSAR data in C-band have a better performance in observing sea ice and land erosion due to its short wavelength and weak penetrability. PolSAR data in L-band, on the other hand, have an advantage in agricultural cover classification due to its long wavelength and strong penetrability. Therefore, it is a promising research topic to make reasonable use of the complementarity between dual-frequency PolSAR data to improve the classification performance.

In the research of dual-frequency PolSAR image classification, the traditional methods are mainly based on statistical models and feature analysis. On the one hand, statistical distribution-based methods provide a good statistical foundation for classification tasks. For example, Dupuis et al. [33] proposed to independently estimate the covariance matrix of each frequency band, and then determine the results based on the sum of Wishart distances of all frequency bands. The essence of this method is to integrate the classification results of each frequency band after processing single-frequency data. In order to integrate dual-frequency data before the Wishart metric processing, Ferro et al. [34] constructed a 6×6 polarization coherence matrix from the scattering matrix of dual-frequency data, and then combined the Wishart metric and maximum likelihood to achieve dual-frequency PolSAR image classification. Gao et al. [35] designed a Wishart mixture model (WMM), which is the weighted sum of multiple distributions. The mixture model realizes interband feature fusion and has stronger descriptive ability than single-frequency data. On the other hand, feature analysis-based methods achieve dual-frequency data fusion by reducing feature redundancy. For example, Liu et al. [36] analyzed the principal components of dual-frequency data by multilinear subspace learning and tensor representation. Yang et al. [37] combined Stein kernel and sparse representation to assume the relationship of dual-frequency data. In addition, De et al. [38] innovatively utilized the Kronecker product to fuse different frequency data, avoiding prior bias caused by the dominant frequency band. Although these traditional algorithms have achieved better results than single-frequency classification algorithms, their classification results are limited because of the weak discriminability of artificial features.

Due to the ability to automatically extract abstract advanced feature representations, DL-based algorithms have achieved better performance than traditional algorithms. Among the DL-based multimodal RS fusion algorithms, Hong et al. [39] created a universal RS basic model based on generative pretrained transformer, which utilizes 3-D token generation to achieve spatial-spectral coupling and adopts a progressive training strategy to adapt to multimodal RS data of different sizes, resolutions, time series, and regions. In [40], a high-resolution domain adaptive network is proposed to solve the domain adaptation problem in multimodal RS data. It preserving the image spatial topology through a parallelled high-to-low resolution fusion model, and reducing the gap caused by the huge differences between different terrain scenes through adversarial learning. Besides, He et al. [41] proposed an adaptive fusion framework, which extracts unimodal features along the spatial and channel dimensions through cross-spatial and cross-channel interaction modules, and establishes a coupled scoring function to describe the dependence relationships to address the differences between modalities. As for PolSAR images, Chen et al. [42] established a dual-frequency PolSAR image classification method based on dynamic neural networks, which has the advantages of fast learning and built-in optimization compared to traditional algorithms. Gadhiya et al. [43] designed an extended optimized Wishart network (e-OWN) based on the mathematical model of Wishart metric, which can accelerate the calculation of Wishart distance and effectively combine dual-frequency PolSAR information. Ahishali et al. [44] adopted 1D-CNN to extract advanced feature representations from single-frequency PolSAR data, and performed dual-frequency feature stacking before the classifier. The experimental results demonstrate that this algorithm improves computational efficiency, and the joint learning of the dual-frequency data significantly upgrades the classification accuracy.

Nevertheless, compared to the mature developed single-frequency PolSAR image classification algorithms, there is still less research on dual-frequency PolSAR image classification. Besides, the current DL-based algorithms usually stack the dual-frequency data and process all features equally, ignoring the complementarity between dual-frequency data. In this case, the

relevant features are not taken seriously while irrelevant features are overly focused, which increases the network training burden and even affects the classification accuracy.

With the aforementioned consideration, a ViT-based model using mixed patch interaction for dual-frequency PolSAR image adaptive fusion classification (PolSAR-MPIformer) is proposed in this article. It mainly includes two parts: the mixed patch interaction (MPI) module and the dual-frequency adaptive fusion (DAF) module. First, the MPI module is used to extract discriminative feature representations by jointly learning global–local spatial information within samples and correlation information among samples. Then, the learned feature representations are fed into the DAF module, which not only makes reasonable use of the complementarity between dual-frequency data by cross-frequency confidence fusion (CFCF) block, but also reduces the speckle noise of PolSAR images through cross-layer attention fusion (CLAF) block. Through the cooperation of the MPI module and the DAF module mentioned above, the goal of improving classification accuracy is achieved. Compared with current state-of-the-art (SOTA) methods, the main increments of the proposed PolSAR-MPIformer are the addition of learning correlation information among samples in the feature extraction process and the use of adaptive fusion strategy to learn dual-frequency data, thereby obtaining more discriminative features and concentrating network computation on favorable frequency band features, achieving reasonable utilization of dual-frequency data. The crucial contributions and novelties of this article are summarized as follows:

1) A novel PolSAR-MPIformer is proposed to improve dual-frequency PolSAR image classification performance. It follows the general framework of ViT and generates an MPI module to extract more discriminative feature representations. In addition, a DAF module is designed as the classifier of PolSAR-MPIformer to fuse dual-frequency data and reduce the influence of speckle noise.
2) An MPI module is designed for feature exaction, which uses patch interaction (PI) intrasample and PI intersample to replace the high-complexity self-attention block in ViT. On the basis of ViT in learning global-spatial information within samples, MPI adds the learning of local-spatial information within samples and correlation information among samples in low network complexity, thus obtaining more discriminative feature representations.
3) A DAF module containing CLAF block and CFCF block is constructed as the classifier of PolSAR-MPIformer. The CLAF block fuses shallow fine-grained features with deep coarse-grained features through the attention mechanism, which reduces the impact of speckle noise while preserving details. Moreover, the CFCF block helps understand the importance of different frequency data for classification tasks through confidence scoring, and applies different weights to different bands accordingly, thereby achieving reasonable utilization of the complementary between dual-frequency PolSAR data.

The article is organized as follows. A brief introduction to PolSAR data is reviewed in Section II. The proposed PolSAR-MPIformer is introduced in Section III. Experimental results and analyses on four real PolSAR datasets are presented in Section IV. Finally, Section V concludes the article.

II. Preliminaries

The PolSAR system obtains descriptions of ground objects by transmitting and receiving electromagnetic waves under various polarization mechanisms. The scattering matrix $\mathbf{S}$ is often used to describe the linear transformation between the transmitted and received electromagnetic waves, and it can be presented as

$$\mathbf{S} = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \tag{1}$$

where H and V denote the horizontal and vertical polarization basis, respectively. Based on the principle of reciprocity (i.e., $S_{HV} = S_{VH}$), we can obtain the Pauli basis $\mathbf{k}$ as follows, and the superscript T indicates transpose operation

$$\mathbf{k} = \begin{bmatrix} S_{HH} & \sqrt{2}S_{HV} & S_{VV} \end{bmatrix}^{\mathrm{T}}. \tag{2}$$

Then, the covariance matrix of $\mathbf{k}$ can be formulated as

$$\mathbf{C} = \mathbf{k} \cdot \mathbf{k}^{H} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{12}^{*} & C_{22} & C_{23} \\ C_{13}^{*} & C_{23}^{*} & C_{33} \end{bmatrix}. \tag{3}$$

The superscript H here represents conjugate transpose operation, and $*$ represents complex conjugate operation. Generally, the internal elements of the covariance matrix $\mathbf{C}$ are selected to form a vector $\mathbf{x}$ as the initial features of each pixel, and $\mathbf{x}$ can be described as $\mathbf{x} = [C_{11}, \mathrm{Re}(C_{12}), \mathrm{Im}(C_{12}), \mathrm{Re}(C_{13}), \mathrm{Im}(C_{13}), C_{22}, \mathrm{Re}(C_{23}), \mathrm{Im}(C_{23}), C_{33}]$, where $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ represent the real and imaginary part, respectively.

For PolSAR image classification tasks, we usually take a window of $s \times s$ centered on pixel to form an image block $(s \times s \times 9)$ as the input sample for deep learning network.

III. Proposed Method

ViT has three core components: the self-attention network, residual connection, and feed forward network (FFN). According to the relevant research of ViT, the key of ViT to achieve excellent performance is not self-attention, but its overall architecture [28]. For this reason, we propose a PolSAR-MPIformer for dual-frequency PolSAR image classification. The structure of the proposed PolSAR-MPIformer is shown in Fig. 1.

Given an input PolSAR image sample $\mathbf{x} \in \mathbb{R}^{s \times s \times 9}$. We first feed $\mathbf{x}$ to several combinations of patch operation and MPI module called stage 0, stage 1, . . . , and stage N for learning layerwise feature representations. Then, these feature representations are fed into the DAF module to achieve dual-frequency data fusion and noise reduction. In the following, we will introduce the designed MPI module and DAF module in detail.

A. Mixed Patch Interaction (MPI) Module

As shown in Fig. 2, the MPI module retains the residual connection and FFN structure in ViT framework, and replaces the high-complexity self-attention network with a combination

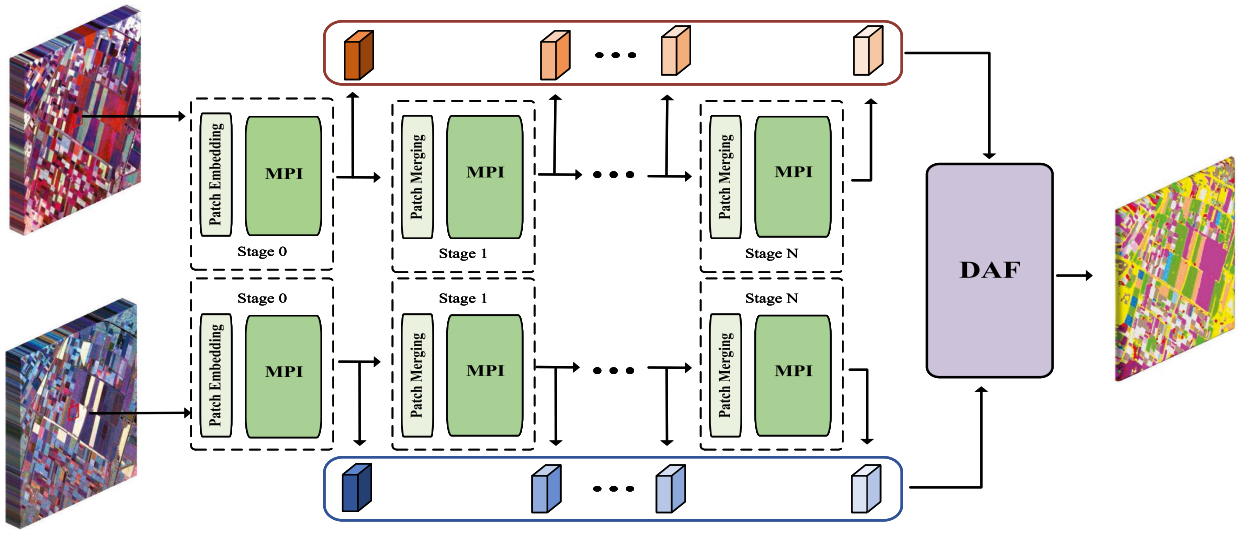

Fig. 1. Overall structure of the proposed PolSAR-MPIformer.

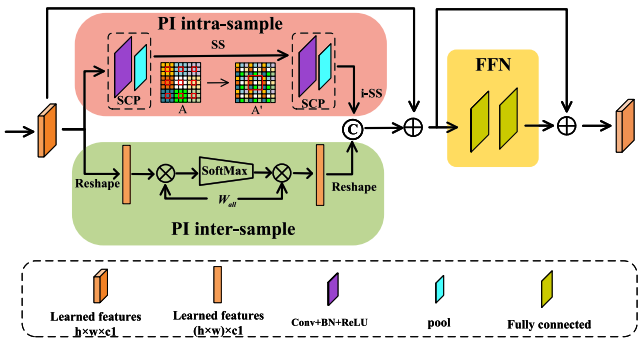

Fig. 2. Architecture of the MPI module.

of PI intrasample and PI intersample. Among them, PI intrasample can be used to learn global–local spatial information within samples, while PI intersample is used to learn correlation information among samples. Below, the PI intrasample and PI intersample are described in detail, respectively.

1) PI Intrasample for Learning Global–Local Spatial Information Within Samples: Due to the powerful global modeling ability, ViT can effectively construct long-range dependence within samples, and has advantages in learning global-spatial information within samples. However, the lack of local-spatial information and the complexity of its structure limits the application of ViT in practice. Hence, by virtue of the advantages of ViT in global-spatial information learning, a PI intrasample method is proposed to jointly extract global and local spatial information.

As a developed image processing network, the convolutional blocks can learn local-spatial information through the local receptive field, shared weights, and spatial subsampling, so it has shift, scale, and distortion invariance. Furthermore, the hierarchical structure of convolutional kernels considers the feature representations with varying degrees of complexity, from simple low-level edges and textures to complex high-level semantic information. Besides, the pooling operation could also complete the learning of local-spatial information to a certain extent, but only using pooling will lose a large amount of image information. Therefore, we adopt a stacked convolution and pooling (SCP) network to learn local-spatial information within samples. This method not only takes the advantage of pooling layer in reducing network parameters, but also combines the advantages of convolutional blocks in preserving local-spatial information.

Apart from local-spatial information over short-range within samples, global-spatial information over long range is also crucial. In ViT, self-attention networks are commonly used to model long-range dependencies. Although this method is effective, the large number of network parameters result in long training time and high training difficulty. Therefore, we propose adding a spatial shuffle (SS) operation between two SCPs to place distant patches in adjacent locations, so that the later SCP can achieve information interaction over long range and learn global-spatial information. Specifically, as shown in the pink area of Fig. 2, the first step of SS is to divide the original sample space A into several nonoverlapping windows. The second step is to move pixels with the same relative position within each window to adjacent positions, ultimately forming a new sample space A'. Both in training and testing follow the same spatial partitioning mode and pixel shuffling mode. In the new sample space A', adjacent patches come from different windows in A. Therefore, when using SCP later, realizing short-range information interaction on A' is equivalent to realizing long-range information interaction on A. Due to the fact that the residual connection is pixelwise addition, we add an inverse SS (i-SS) after the second SCP to ensure pixel alignment between the processed sample and the original sample.

Compared with self-attention in ViT, the PI intrasample method based on SCP and SS can learn global–local spatial information and has lower network complexity. Therefore, as shown in the pink area in Fig. 2, for PolSAR image classification tasks, we adopt the PI intrasample method to learn global–local spatial information within samples.

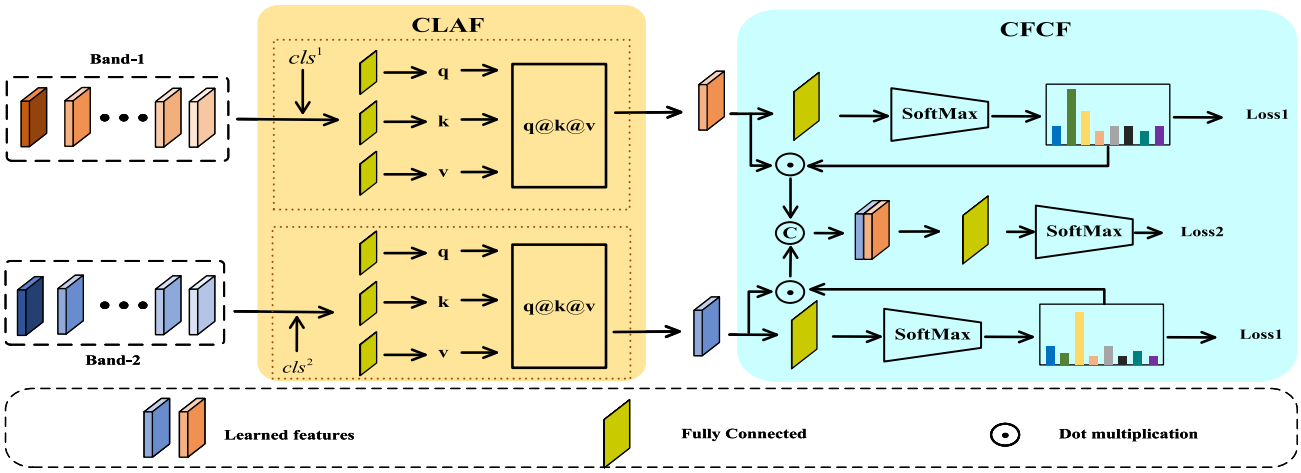

Fig. 3. Architecture of the DAF module.

2) PI Intersample for Learning Correlation Information Among Samples: For the whole PolSAR image, the same kind of pixels may appear at a far distance. When samples are generated by window segmentation, similar pixels may appear in different samples. Therefore, it is necessary to consider the correlation information among samples. Below, we will introduce a PI intersample method to achieve feature learning among samples.

Self-attention mechanism is a common algorithm for information interaction, and is often used for feature learning among patches within the sample. When extending the self-attention mechanism to information interaction among samples, it is necessary to first measure the correlation between the input sample and all other samples, and then use the obtained correlation matrix to apply different weights to all samples. The mathematical formulation of this operation is

$$\mathbf{x_{out}} = \mathrm{SoftMax}(\mathbf{x_{in}} \times \mathbf{W_{all}^{T}}) \times \mathbf{W_{all}} \tag{4}$$

where $\mathbf{x_{in}}$ and $\mathbf{x_{out}}$ represent the sample information of input and output, respectively, $\mathbf{W_{all}}$ represents the sample information of the entire PolSAR image. This concise formula inspires a network architecture to realize automatic information interaction among samples. As shown in the green area in Fig. 2, instead of stacking all samples to obtain correlation matrix and update sample information, we abstract the matrix $\mathbf{W_{all}}$ as a network parameter that can gradually fit the sample information of whole PolSAR image through the training process, thus solving the problem of huge calculation caused by measuring the correlation among samples one by one.

After combining the correlation information intersample with the spatial information intrasample, a more complete feature representation can be obtained, which provides a more favorable feature basis for subsequent classifier.

B. Dual-Frequency Adaptive Fusion (DAF) Module

Through the MPI module, the shallow feature representations and deep feature representations of dual-frequency PolSAR image can be obtained. Because the shallow fine-grained feature representations contain detailed information and the deep coarse-grained feature representations contain semantic information, we propose a CLAF block to make full use of these two kinds of features. In view of the different sensitivities of dual-frequency data to ground objects, we propose a dual-frequency data fusion block named CFCF, so as to reasonably utilize the complementarity of dual-frequency data and produce more accurate classification results. The specific descriptions are as follows.

1) Cross Layer Attention Fusion (CLAF): The normal DL-based methods only input the feature representations from the last layer into the classifier, but ignore the spatial details contained in the shallow feature representations and the correlation information across different layers. In order to make the best of shallow feature representations and deep feature representations obtained by MPI modules, the CLAF block is proposed.

Specifically, as shown in the yellow area of Fig. 3, the CLAF enables shallow feature representations and deep feature representations to learn from each other through self-attention network. Due to the same processing method for all the frequency bands, we will take Band-1 as an example for a detailed introduction.

First, the multilayer feature representations are stacked as

$$\mathbf{X_{in}^{1}} = [\mathbf{cls^{1}}, \mathbf{x_{1}^{1}}, \ldots, \mathbf{x_{n}^{1}}, \ldots, \mathbf{x_{N}^{1}}] \tag{5}$$

where N denotes the number of network layers, $\mathbf{x_{n}^{1}}$ indicates the nth layer feature representations of Band-1 data. Inspired by transformer, a feature vector $\mathbf{cls^{1}}$ is embedded in the first place of $\mathbf{X_{in}^{1}}$. Next, the self-attention mechanism is used to construct a query vector $\mathbf{q}$, a key vector $\mathbf{k}$, and a value vector $\mathbf{v}$. The relationship across layers can be found through $\mathbf{q}$ and $\mathbf{k}$. Then, the value vector $\mathbf{v}$ is weighted by the obtained coefficient matrix to generate the fusion features. The specific mathematical expression is

$$\mathbf{X_{out}^{1}} = \mathrm{softmax}(\mathbf{q}\mathbf{k}^{T})\mathbf{v}. \tag{6}$$

The dimensions of $\mathbf{X_{out}^{1}}$ are the same as $\mathbf{X_{in}^{1}}$, and we choose the first vector of $\mathbf{X_{out}^{1}}$ as the cross layer feature representations for output.

After the above operations, the CLAF block could transfer the semantic information of deep feature representations to shallow layers, thus reducing speckle noise in shallow fine-grained feature representations. In addition, it also transfers the detailed information of shallow feature representations to deep layers, so that the detailed information can be retained in deep coarse-grained feature representations. Although the CLAF block uses the self-attention mechanism, it does not bring in a large number of network parameters and excessive computational complexity. There are two main reasons, on the one hand, the computation complexity of self-attention is $O(n^2)$, where n is the number of image patches in ViT and the number of MPI modules in the proposed algorithm. The number of MPI modules is much less than the number of normal image patches, so the computation complexity of self-attention in CLAF is lower than that in ViT. On the other hand, the CLAF block only appear once after the feature extraction process and do not add too much computational complexity to the entire network. To summarize, through the CLAF block mentioned above, the impact of speckle noise on classification results can be reduced, and the detailed information can be kept as much as possible.

2) Cross Frequency Confidence Fusion (CFCF): Due to the different sensitivities of dual-frequency PolSAR data to ground objects, their classification performance are different. In order to fully utilize the differences in dual-frequency data and improve the classification performance, we propose a CFCF block that utilizes the complementary of dual-frequency data through confidence scoring.

By setting different classifiers for different frequency data, two different classification results can be obtained. Specifically, we use the combination of fully connected (FC) layer and softmax operation as classifiers, and the number of output neurons is K, which is the number of categories of ground objects. The output of Band-1 classifier and Band-2 classifier can be expressed as $\mathbf{out1}$ and $\mathbf{out2}$, respectively:

$$\begin{aligned}\mathbf{out^1} &= \text{SoftMax}(FC(\mathbf{feature^1})) \\ &= [y_1^1, \ldots, y_k^1, \ldots, y_K^1] \end{aligned} \tag{7}$$

$$\begin{aligned}\mathbf{out^2} &= \text{SoftMax}(FC(\mathbf{feature^2})) \\ &= [y_1^2, \ldots, y_k^2, \ldots, y_K^2]. \end{aligned} \tag{8}$$

In the above equations, $\mathbf{feature^1}$ and $\mathbf{feature^2}$ represent the feature representations of Band-1 and Band-2 obtained by CLAF block, respectively, y_k^1 and y_k^2 represent the output of the kth neuron in Band-1 classifier and Band-2 classifier.

The negative logarithmic likelihood (NLL) loss caused by the outputs and the one-hot coded labels $\mathbf{y}$ is propagated back, and the classifiers of two frequency bands can be trained, respectively. The loss function of this part is presented as follows:

$$\text{loss}_1 = -\mathbf{y} \log \mathbf{out^1} - \mathbf{y} \log \mathbf{out^2}. \tag{9}$$

In the prediction process, the index of the maximum value in output vector is generally selected as the classification result, but the numerical characteristics of $\mathbf{out^1}$ and $\mathbf{out^2}$ are ignored. Taking $\mathbf{out^1}$ as an example, the physical meaning of each item y_k^1 is the probability that the sample is classified into the kth class on Band-1 data. Similarly, the physical meaning of y_k^2 is the probability that the sample is classified into kth class on Band-2 data. Therefore, although the index of the largest item in $\mathbf{out^1}$ and $\mathbf{out^2}$ should be the same, the numerical results of $\mathbf{out^1}$ and $\mathbf{out^2}$ are different.

The purpose of network training is to make the predicted label as close as possible to the real label, that is, the target items of $\mathbf{out^1}$ and $\mathbf{out^2}$ should be close to 1, and other items should be close to 0. Therefore, the greater the gap between the largest term and the second largest term, the closer the output is to the ideal state, indicating that this group of feature representations has stronger classification ability. Based on this, we convert the difference value between the largest term and the second largest term as the classification confidence of each band. By applying different weights to different frequency bands based on the classification confidence, the fused feature representations $\mathbf{feature}$ can be obtained as

$$\alpha = y_{k1}^1 - y_{k2}^1 \tag{10}$$

$$\beta = y_{k1}^2 - y_{k2}^2 \tag{11}$$

$$\mathbf{feature} = [\alpha * \mathbf{feature^1}, \beta * \mathbf{feature^2}] \tag{12}$$

where y_{k1}^1 and y_{k2}^1 represent the largest term and second largest term of $\mathbf{out^1}$, y_{k1}^2 and y_{k2}^2 represent the largest term and second largest term of $\mathbf{out.^2}$

Then the fused feature representation $\mathbf{feature}$ is fed into a classifier to get $\mathbf{out}$, the generated NLL loss of this part is defined as

$$\text{loss}_2 = -\mathbf{y} \log \mathbf{out}. \tag{13}$$

To sum up, the final overall loss function can be written as

$$\text{loss} = \text{loss}_2 + \lambda \times \text{loss}_1 \tag{14}$$

where λ is the weight coefficient. By combining loss_1 and loss_2 for backpropagation, the best fusion result can be obtained while ensuring the appropriate confidence score for each frequency band as accurate as possible.

The above confidence-based dual-frequency feature fusion algorithm can fully learn the rich information contained in dual-frequency data, and make reasonable use of their complementarity to improve the classification accuracy.

IV. Experimental Result and Analysis

In this section, we perform experiments on four measured dual-frequency PolSAR datasets to evaluate the performance of the proposed PolSAR-MPIformer. The experimental setup, ablation studies, and performance comparison are introduced below.

A. Experimental Setup

1) Dataset: Three sets of real multifrequency PolSAR images are used for the following experiments. The ground truth maps of each dataset are manually annotated based on the corresponding optical images on Google Earth.

The first dataset has C-, L-, and P-band PolSAR data obtained by the NASA/JPL AIRSAR system in the Flevoland region, with

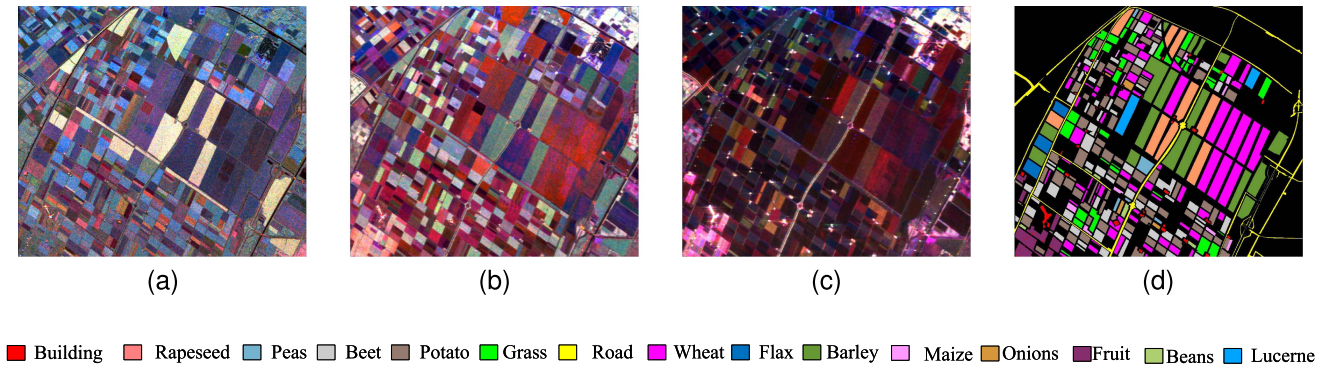

Fig. 4. Flevoland dataset. (a) C-band. (b) L-band. (c) P-band. (d) Ground Truth.

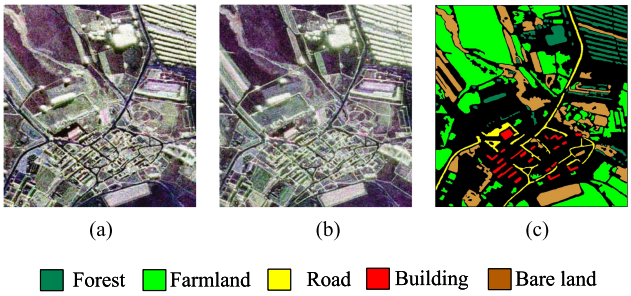

Fig. 5. Hebei dataset. (a) S-band. (b) L-band. (c) Ground Truth.

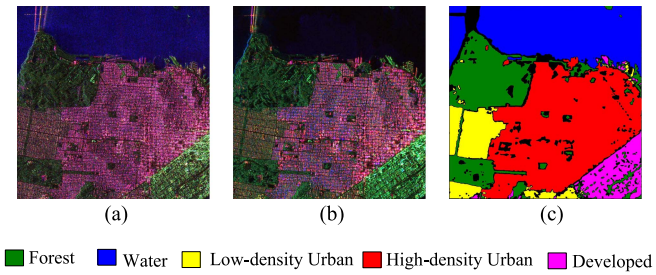

Fig. 6. SanFrancisco dataset. (a) C-band. (b) L-band. (c) Ground Truth.

a size of 1079×1024 and containing 15 types of ground objects. Fig. 4(a) to (c) shows the Pauli decomposition images of each band and Fig. 4(d) shows the ground truth map.

The second dataset contains S- and L-band PolSAR data obtained by the airborne system over the scene of Hebei in 2021. It has 1005×962 pixels and five types of ground objects. The Pauli decomposition images of S- and L-band data are shown in Fig. 5(a) and (b). The referenced ground truth image is shown in Fig. 5(c).

The third dataset has C-band PolSAR data collected by the GF-3 system and L-band PolSAR data collected by the ALOS system. It displays the terrain situation of San Francisco Bay, and mainly includes five kinds of ground objects. The image size is 1161×1161 and the Pauli decomposition images of C- and L-band data are presented in Fig. 6(a) and (b), respectively. The ground truth image is presented in Fig. 6(c).

In order to apply the above measured PolSAR data to this study, we constructed four sets of dual-frequency PolSAR datasets: the *Flevoland_CL* contains C- and L-band PolSAR data from Flevoland dataset, the *Flevoland_CP* contains C- and P-band PolSAR data from Flevoland dataset, the *Hebei_SL* contains S- and L-band PolSAR data from Hebei dataset, and the *SanFrancisco_CL* contains C- and L-band PolSAR data from SanFrancisco dataset. In addition, the RS images collected from airborne or satellite equipment suffered from various variabilities [45], [46], PolSAR data always take surface scattering, double-bounce scattering, volume scattering, helix scattering as the endmember matrix (or dictionary) of multiple-component scattering model. In this study, we only focus on observing interclass discrimination and intraclass similarity, without considering the detailed scattering characteristics in each type of terrain.

2) Implementation Details: In the following experiments, we select 13×13 as the window size to construct samples. Due to the large number of categories in the Flevoland dataset, we randomly select 200 samples from each class as the training sample set for *Flevoland_CL* and *Flevoland_CP*. The Hebei dataset and SanFrancisco dataset contain five types of ground objects, which is a relatively small number of categories. Therefore, 100 samples are selected from each class as the training sample set for *Hebei_SL* and *SanFrancisco_CL*.

In the training process, we set the epoch as 200 and the optimizer as Adam, the basic learning rate is $1e{-}3$, which is decayed by 0.9 every 50 epochs. All the experiments are performed on the Hewlett-Packard (HP)-Z840 Workstation with Nvidia GeForce RTX 1080 GPU, 64-GB RAM, and Windows 10 operating system. In order to avoid the influence of randomness, we run all the algorithms independently for 10 times in PyTorch environment and take the average value as the final result.

As for evaluation indicators, we take the accuracies of each category, overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) to measure the effectiveness of the algorithms.

B. Ablation Studies

The proposed PolSAR-MPIformer mainly includes the MPI module and DAF module. To verify the effectiveness of each module, we conduct ablation experiments. Moreover, we design experiments on single- and dual-frequency data separately to

TABLE I
ABLATION STUDY OF THE MPI MODULE

Method	Flevoland_CL	Flevoland_CP	Hebei_SL	SanFrancisco_CL
	OA(%)/AA(%)/$\kappa\times$ 100			
ViT	97.28/97.63/96.87	96.57/97.04/96.06	95.38/95.29/92.85	98.70/98.38/98.24
PI intra-sample	97.74/98.14/97.40	97.06/97.44/96.67	96.77/95.74/94.99	98.89/98.62/98.55
PI inter-sample	97.45/97.76/97.07	96.24/96.74/95.68	96.21/95.66/94.12	98.85/98.68/98.42
MPI	**97.96/98.23/97.66**	**97.25/97.65/96.84**	**97.13/96.06/95.52**	**98.96/98.74/98.57**

The bold values represent the best results.

verify the effectiveness of dual-frequency data fusion in classification tasks. Due to the fact that the loss function contains two terms during the training process, we also conduct experiments on the weight coefficient λ.

1) Effect of MPI Module: In this section, we use four models for feature extraction, respectively, to verify the effectiveness of MPI module. These four models are ViT model, PI intrasample model, PI intersample model, and MPI model. To avoid the impact of DAF module on experimental results, a simple combination of FC layer and softmax operation is used as the classifier.

As shown in Table I, the classification results of PI intrasample are superior to ViT model in all the four datasets, which is related to the sampling method of PolSAR image classification task. Due to the small window in the sampling process, it is very wasteful for ViT to use a complex network to conduct long-range information interaction within samples. The complexity of the model leads to the increased training difficulty and network overfitting. Besides, ViT only focuses on extracting global-spatial information while ignoring local-spatial information. These reasons collectively result in the accuracy of ViT being lower than the accuracy of PI intrasample.

In addition, due to the fact that the PI intrasample only achieves information interaction within samples but ignores the correlation information among samples, a PI intersample is proposed to supplement the feature extraction process. It can be seen from Table I that the accuracy of PI intersample is lower than that of PI intrasample. This is because the correlation across different samples is less strong than the correlation within samples, and the feature representations obtained within samples are naturally more discriminative than that obtained across samples. Nevertheless, the correlation information among samples still contains some important classification information. The classification results of MPI in Table I clearly indicate that the feature extraction algorithm combining PI intrasample and PI intersample can obtain more discriminative feature representations.

The above experiments have demonstrated that the proposed MPI module can fully utilize the sample information to learn discriminative feature representations, and provide a good foundation for subsequent steps.

2) Effect of DAF Module: The DAF module has two important blocks, namely, CLAF block and CFCF block. Therefore, we designed four models to prove the effectiveness of DAF module. The first ablation model directly uses FC layer as the classifier, the second ablation model uses CLAF before the FC layer, the third ablation model uses CFCF before the FC layer, and the fourth ablation model uses both CLAF and CFCF before the FC layer. The above four models are all based on the MPI module for feature extraction.

From Table II, we can observe that the classifier with CLAF or CFCF is better than the classifier with only FC layer. The CLAF could utilize the combination of shallow and deep feature representations based on the self-attention mechanism, thereby reducing the impact of speckle noise on experimental results while preserving details. And CFCF is a fusion method for dual-frequency PolSAR data. Based on the different sensitivities of dual-frequency data on terrain classification, the CFCF uses the complementarity between dual-frequency data to improve the classification accuracy. The CLAF and CFCF have improved the classifier from different perspectives, and both have made significant progress.

The DAF module suggests applying CLAF and CFCF to the classifier simultaneously. As can be seen from Table II, the DAF module can not only utilize the advantages of CLAF to learn hierarchical information to reduce speckle noise, but also utilize the advantages of CFCF to learn complementary information between dual-frequency data to improve classification results.

3) Effect of Dual-Frequency Data Fusion: Due to the fact that single-frequency PolSAR image classification does not require the CFCF block, we remove it from the DAF module in the single-frequency experiments. The color coding in tSNE maps below is the same as the ground truth maps.

The classification results of the Flevoland dataset are shown in Table III, and the tSNE maps are shown in Fig. 7. As can be seen from Table III, L-band performs best in three frequency bands, and maintains the highest classification accuracy in most categories. This is because the wavelength and resolution of L-band are suitable for detecting forestry and agricultural ground objects. As can be seen from Fig. 7, C-band has more compact features in Oats and Beans [as shown in the elliptical box and rectangular box in Fig. 7(a)]. In addition, compared with C-band, P-band has less overlap area in Wheat and Barley [as shown in the elliptical box in Fig. 7(c)]. Although L-band is the best of the three bands in terms of accuracy, there are still problems of intermixing across categories and insufficient compactness within categories in the tSNE map. In Fig. 7(d), it is obvious that after the dual-frequency fusion, the Wheat and Barley in C+L are more isolated than those in C-band [as shown in the elliptical box in Fig. 7(d)], the Maize is more compact than those in the single-frequency situation [as shown in the rectangular box

TABLE II
ABLATION STUDY OF THE DAF MODULE

MPI	DAF		Flevoland_CL	Flevoland_CP	Hebei_SL	SanFrancisco_CL
	CLAF	CFCF	OA(%)/AA(%)/$\kappa\times$ 100			
✔	✘	✘	97.96/98.23/97.66	97.25/97.65/96.84	97.13/96.06/95.52	98.96/98.74/98.57
✔	✔	✘	98.09/98.21/97.80	97.43/97.76/97.04	97.45/95.59/95.99	98.98/98.06/98.60
✔	✘	✔	98.16/98.36/97.88	97.68/97.79/97.33	97.90/96.69/96.71	99.17/98.85/98.86
✔	✔	✔	**98.34/98.45/98.09**	**97.83/98.04/97.50**	**98.45/96.74/97.57**	**99.31/98.86/99.05**

The bold values represent the best results.

TABLE III
CLASSIFICATION PERFORMANCE OF FLEVOLAND DATASET WITH DIFFERENT BANDS

Band	**C**	**L**	**P**	**C+L**	**C+P**	**C+L+P**
Grass	93.12	91.41	87.32	**97.03**	96.92	96.44
Flax	99.86	99.86	96.90	99.88	99.91	**99.94**
Potato	96.77	98.35	91.87	98.87	98.94	**99.19**
Wheat	93.92	97.27	97.08	98.77	98.86	**98.92**
Rapeseed	99.54	**99.84**	99.19	99.80	99.66	99.72
Beet	86.04	95.35	83.06	98.71	96.02	**99.02**
Barley	93.74	98.96	98.29	**99.46**	98.92	99.30
Peas	99.40	98.60	97.91	**99.88**	99.52	99.78
Maize	96.44	**99.72**	91.95	99.63	99.35	99.57
Beans	99.04	99.85	98.89	**99.60**	99.09	98.84
Fruit	92.03	96.53	96.17	**97.89**	96.66	96.95
Onions	99.64	98.23	97.64	99.68	**99.91**	99.86
Lucerne	97.84	99.65	95.36	99.98	99.77	**100.00**
Building	92.72	93.87	95.26	95.33	96.01	**97.27**
Road	85.29	86.38	81.92	92.21	91.01	**92.62**
OA(%)	93.36	96.66	93.06	98.34	97.83	**98.41**
AA(%)	94.91	97.04	93.92	98.45	98.04	**98.49**
$\kappa\times100$	92.38	96.16	92.03	98.09	97.50	**98.17**

The bold values represent the best results.

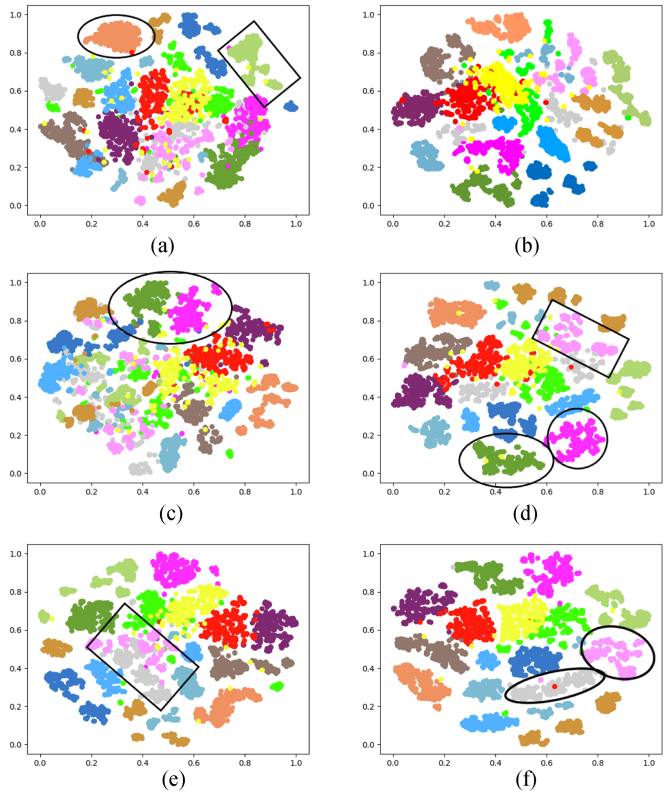

Fig. 7. Comparison tSNE Maps of Flevoland Dataset. (a) C-band. (b) L-band. (c) P-band. (d) C+L. (e) C+P (f) C+L+P.

in Fig. 7(d)]. As for C+P, although the separability of Beat and Maize is not yet strong, there has been a significant improvement compared to the results of single-frequency [as shown in the rectangular box in Fig. 7(e)].

Moreover, we extend our algorithm to the trifrequency PolSAR data. The experimental result is shown in Fig. 7(f), and it is clear that the separability across categories and the compactness within categories are both strong. For example, the confusion between Beat and Maize is significantly less than other results [as shown in the elliptical boxes in Fig. 7(f)]. From Table III, we can also see that the classification accuracy of trifrequency data is higher than that of single-frequency data and dual-frequency data. However, due to the high hardware requirements for obtaining and processing trifrequency PolSAR data, there are very few relevant measured datasets currently. Moreover, the fusion of dual-frequency data has effectively improved classification accuracy, the accuracy of trifrequency data fusion is not significantly improved compared to dual-frequency data. According to the above analysis, we mainly focus on the dual-frequency dataset (*Flevoland_ CL* and *Flevoland_CP*) in subsequent experiments.

TABLE IV
CLASSIFICATION PERFORMANCE OF HEBEI DATASET WITH DIFFERENT BANDS

Band	**S**	**L**	**S+L**
Forest	97.99	97.50	**98.65**
Farmland	98.12	**99.83**	99.05
Road	93.49	37.58	**91.84**
Building	86.15	87.22	**94.81**
Bare land	93.05	94.95	**99.09**
OA(%)	96.16	94.15	**98.45**
AA(%)	93.76	83.41	**96.69**
$\kappa\times100$	93.98	90.52	**97.57**

The bold values represent the best results.

The experimental results on two frequency data of the *Hebei_SL* are shown in Table IV. It can be seen that the classification accuracy of S-band in Road is higher than that of L-band, and the classification accuracy of L-band in Farmland is higher than that of S-band. Moreover, the tSNE maps in Fig. 8 show that the

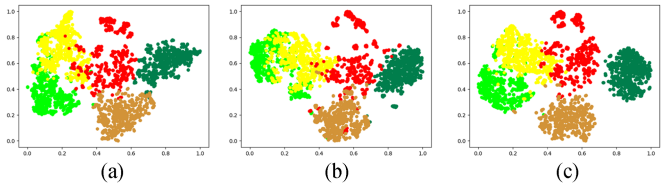

Fig. 8. Comparison tSNE Maps of Hebei Dataset. (a) S-band. (b) L-band. (c) S+L.

TABLE V
CLASSIFICATION PERFORMANCE OF SANFRANCISCO DATASET WITH DIFFERENT BANDS

Band	C	L	C+L
Forest	95.25	95.94	**98.55**
Water	97.30	99.85	**99.86**
High-density urban	98.50	96.83	**99.73**
Low-density urban	97.94	94.83	**98.64**
Developed	95.97	94.22	**97.46**
OA(%)	97.37	97.16	**99.31**
AA(%)	96.99	96.33	**98.85**
$\kappa\times100$	96.38	96.09	**99.05**

The bold values represent the best results.

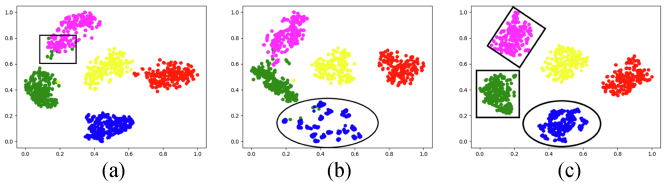

Fig. 9. Comparison tSNE Maps of SanFrancisco Dataset. (a) C-band. (b) L-band. (c) C+L.

Road and Farmland are severely mixed under single-frequency data. After the dual-frequency data fusion, although there is still class overlap, the results have been greatly improved.

The experimental results on *SanFrancisco_CL* are shown in Table V and Fig. 9. As shown in the tSNE maps of C-band and L-band, the Water of C-band is more compact, while the Water of L-band is more dispersed [as shown in the elliptical box in Fig. 9(b)]. However, the misclassification of forest and developed of C-band is more serious than that of L-band [as shown in the rectangular box in Fig. 9(a)]. After the fusion of dual-frequency data, it can be seen that the separability across categories and the compactness within categories have been improved. Table V also shows that the accuracy of dual-frequency data on each class is better than that of single-frequency data.

4) Effect of λ*:* In this section, we conduct experiments under different tradeoff parameter λ, and present the experimental results in Fig. 10. For *Flevoland_CL*, high classification accuracy can be achieved when λ is between 0.2 and 0.7, with the optimal λ being 0.5. For *Flevoland_CP* and *Hebei_SL*, OA, AA, and κ reach the peak when $\lambda = 0.2$ and $\lambda = 0.9$, respectively. As the accuracy of *SanFrancisco_CL* is already very high, almost reaching over 98.5%, small fluctuations in accuracy are allowed. As shown in Fig. 10(d), the general trend of classification results is increased earlier and decreased later, and reaching the optimal value at $\lambda = 0.7$ approximately.

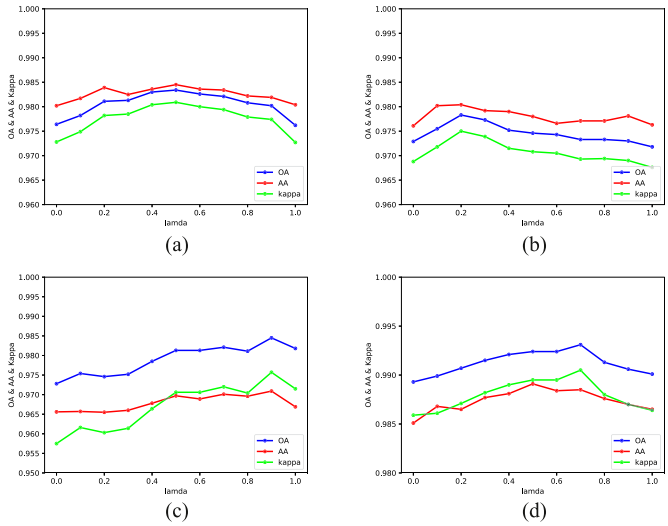

Fig. 10. Effects of the tradeoff parameter λ on classification performance. (a) Flevoland_CL. (b) Flevoland_CP. (c) Hebei_SL. (d) SanFrancisco_CL.

Based on the above experimental results, we can see that the second term of the loss function has indeed contributed to improving the classification accuracy.

C. Performance Comparison

In this section, we compare the proposed PolSAR-MPIformer with several state-of-the-art dual-frequency data fusion algorithms. On the one hand, WMM [35], online multiview deep forest (OMDF) [47], and Kronecker product (KP) [38] are compared as non-DL fusion methods to demonstrate the superiority of neural networks in feature extraction. On the other hand, five DL-based fusion methods, including two-branch CNN (t-CNN) [48], feature intersecting learning-based CNN (FIL-CNN) [49], cross channel reconstruction network (CCR-Net) [50], global–local transformer (GLT) [51], and multimodal fusion network (MFNet) [52], are chosen for comparison. Specifically, t-CNN, FIL-NN, CCR-Net, and MFNet are all based on CNN for feature extraction, while GLT is based on CNN and transformer for feature extraction. In addition, t-CNN, CCR-Net, and GLT achieve multimodal RS image fusion by feature concatenation, while FIL-NN and CCR-Net achieve multimodal RS image fusion by attention strategy. By conducting comparative experiments on these fusion methods, we aim to verify the effectiveness of the proposed PolSAR-MPIformer in feature extraction and dual-frequency data fusion. For a fair comparison, we optimize the parameters of all models on the same hardware equipment and the same training samples.

1) Results on Flevoland_CL*:* The classification accuracy of each category, OA, AA, and κ on *Flevoland_CL* are represented in Table VI. It can be seen that the proposed PolSAR-MPIformer achieves the optimal classification accuracy on OA, AA, and κ, and has the highest accuracy in most categories. The visual classification results on *Flevoland_CL* are shown in Fig. 11. Compared with the DL-based algorithms, the classification map of WMM is severely affected by speckle noise [highlighted by the rectangular and elliptical boxes in Fig. 11(b)]. This is because

TABLE VI
COMPARISON CLASSIFICATION ACCURACY ON *FLEVOLAND_CL*

Method	WMM	OMDF	KP	t-CNN	FIL-CNN	CCR-Net	GLT	MFNet	Ours
Grass	35.73	91.07	88.31	94.52	93.42	96.23	93.11	96.94	**97.03**
Flax	95.42	99.83	99.85	99.71	99.88	**99.93**	99.91	99.84	99.88
Potato	96.33	96.33	98.13	90.81	98.86	**99.29**	97.70	99.09	98.87
Wheat	91.40	95.88	96.75	95.16	98.46	97.53	97.11	98.17	**98.77**
Rapeseed	**99.96**	99.94	99.82	99.96	99.84	99.86	99.10	99.79	99.80
Beet	32.95	76.05	88.53	97.50	98.02	97.48	95.81	96.99	**98.71**
Barley	97.57	97.68	98.45	97.17	99.22	**99.59**	99.19	99.34	99.46
Peas	65.82	**100.00**	98.89	**100.00**	99.95	99.86	99.28	99.81	99.88
Maize	76.49	97.51	97.61	92.60	99.31	99.70	98.89	**99.65**	99.63
Beans	98.44	98.59	95.86	99.29	98.03	99.60	98.59	99.04	**99.60**
Fruit	97.89	91.72	95.46	93.08	96.69	97.34	97.26	97.32	**97.89**
Onions	95.47	99.00	99.14	99.23	**99.68**	99.64	98.23	99.55	**99.68**
Lucerne	24.83	99.84	99.16	99.91	**99.98**	99.95	99.95	**99.98**	**99.98**
Building	23.43	86.68	91.64	**97.66**	95.41	94.91	93.36	95.15	95.33
Road	41.89	91.63	85.40	89.79	92.12	91.42	86.31	90.34	**92.21**
OA(%)	77.42	93.73	95.15	95.22	97.85	97.89	96.56	97.86	**98.34**
AA(%)	71.57	94.78	95.53	96.43	97.92	98.15	96.92	98.04	**98.45**
$\kappa\times100$	74.10	92.82	94.43	94.52	97.53	97.57	96.04	97.54	**98.09**

The bold values represent the best results.

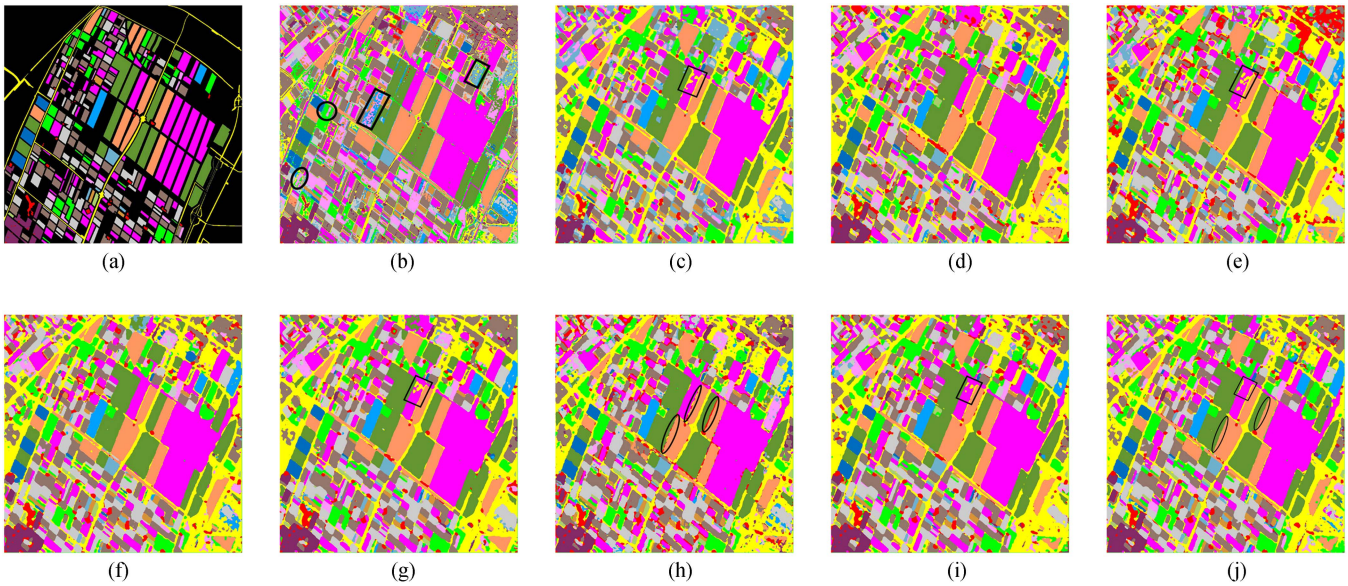

Fig. 11. Comparison classification maps on *Flevoland_CL*. (a) Ground Truth. (b) WMM. (c) OMDF. (d) KP. (e) t-CNN. (f) FIL-Net. (g) CCR-Net. (h) GLT. (i) MFNet. (j) Ours.

WMM, as a traditional pixel-level classification method, only focuses on the target pixel in isolation and does not consider contextual neighborhood information, resulting in many noise points in homogeneous region. In the experiments on other datasets, we can also see that the classification maps of WMM do have significant noise. For t-CNN, CCR-Net, and MFNet with high classification accuracy, there are still some obvious misclassifications, which are highlighted by the rectangular boxes in each figure. That is because the convolutional kernels only focus on the local-spatial information and cannot consider the long-range information. As for GLT, it uses transformer and CNN for global and local spatial information learning, respectively, but ignores the correlation information among samples, which may lose some discriminative information and lead to misclassification. The result of GLT is shown in Fig. 11(h), there are indeed some severe misclassifications highlighted by elliptical boxes. Among all the algorithms, FIL-CNN and the proposed PolSAR-MPIformer are the best in visual results, and there is less noise in consistent areas. As shown in Table VI, FIL-CNN can achieve the highest accuracy in Onions and Lucerne. However, its accuracy in Grass is average, while the proposed PolSAR-MPIformer can achieve 97% in Grass, which is the highest among all algorithms. Moreover, the classification accuracy of the proposed PolSAR-MPIformer on Onions and Lucerne is the same as FIL-CNN, both of which are the highest among these algorithms. In summary, compared with several state-of-the-art dual-frequency classification algorithms, the proposed PolSAR-MPIformer achieves high classification accuracy while maintaining spatial consistency on *Flevoland_CL*.

2) Results on Flevoland_CP*:* Eight related classification algorithms are used for comparative experiments on *Flevoland_CP*, and the results are shown in Table VII and

TABLE VII
COMPARISON CLASSIFICATION ACCURACY ON *FLEVOLAND_CP*

Method	WMM	OMDF	KP	t-CNN	FIL-CNN	CCR-Net	GLT	MFNet	Ours
Grass	28.13	**97.75**	87.91	92.29	95.89	95.97	94.47	96.41	96.92
Flax	90.09	99.80	99.64	99.84	99.83	99.80	**99.94**	99.88	99.91
Potato	87.80	95.57	98.22	89.73	98.64	98.29	97.35	98.52	**98.94**
Wheat	95.39	87.35	96.36	98.53	**98.87**	98.28	96.32	98.56	98.86
Rapeseed	99.83	**99.93**	99.67	99.91	99.83	99.56	99.48	99.76	99.66
Beet	38.38	**97.17**	82.66	83.08	93.53	93.99	88.29	94.02	96.02
Barley	95.04	98.32	96.55	94.39	98.33	98.81	98.02	98.82	**98.92**
Peas	66.04	79.81	97.06	**99.95**	99.86	99.83	98.63	99.11	99.52
Maize	63.35	52.16	94.95	87.55	99.24	**99.72**	98.57	98.57	99.35
Beans	98.23	**99.95**	94.45	98.18	99.14	98.89	98.84	98.49	99.09
Fruit	88.34	92.35	96.00	96.48	96.09	96.15	96.80	96.53	**96.66**
Onions	78.06	**100.00**	99.05	98.69	99.55	99.27	98.55	99.55	99.91
Lucerne	20.00	98.11	99.06	**99.87**	99.70	99.41	99.46	99.79	99.77
Building	43.14	**97.82**	92.28	94.82	96.23	96.96	91.69	95.02	96.01
Road	34.59	76.93	83.94	**91.22**	90.60	90.68	83.15	87.59	91.01
OA(%)	75.08	92.93	93.93	93.70	97.31	97.24	95.13	97.13	**97.83**
AA(%)	67.09	91.53	94.52	94.97	97.69	97.71	95.97	97.37	**98.04**
$\kappa\times100$	71.58	91.90	93.03	92.78	96.90	96.83	94.41	96.70	**97.50**

The bold values represent the best results.

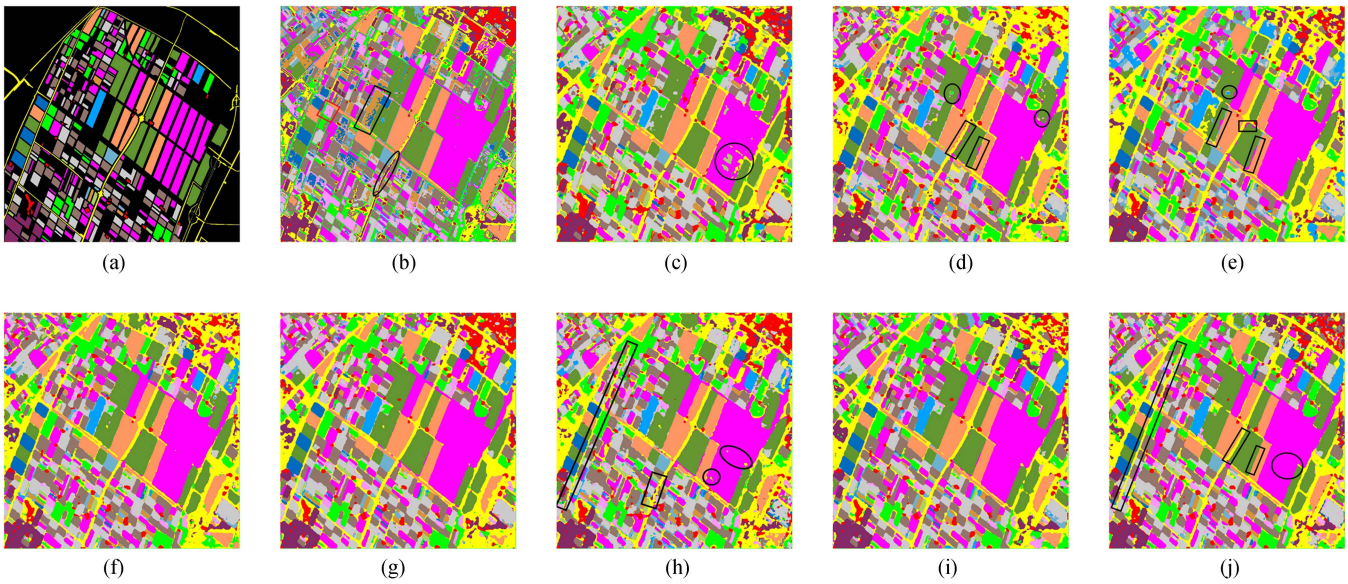

Fig. 12. Comparison classification maps on *Flevoland_CP*. (a) Ground Truth. (b) WMM. (c) OMDF. (d) KP. (e) t-CNN. (f) FIL-Net. (g) CCR-Net. (h) GLT. (i) MFNet. (j) Ours.

Fig. 12. In Fig. 12(b), WMM misclassifies Lucerne as Onions (as shown in the rectangular box) and Road as Fruit (as shown in the elliptical box). The classification result of OMDF is shown in Fig. 12(c), there is an obvious misclassification in the middle of the classification map (as shown in the rectangular box). For Fig. 12(d), KP combines the dual-frequency PolSAR data through mathematical methods. Although KP is not a DL-based fusion algorithm, it utilizes the neural network for subsequent feature learning, resulting in better classification results than WMM. For Fig. 12(e), the t-CNN constructed two branches to learn two frequency data independently, and stacked them in the last layer for data fusion. The redundancy caused by feature accumulation will increase the training burden and affect the classification results. In Fig. 12(d) and (e), the rectangular boxes indicate significant misclassification at the boundary pixels. In addition, some consistent areas with internal noise are also marked by the elliptical box in Fig. 12(d) and (e). In Fig. 12(h), we can see that some road areas are misclassified as other categories (highlighted by rectangles), and some wheat areas misclassified as road (highlighted by ovals). Relatively speaking, the classification results of Fig. 12(f), (g), (i), and (j) are smooth, and the noise in consistency region is low. Combined with the classification results in Table VII, we can see that the FIL-CNN, CCR-Net,

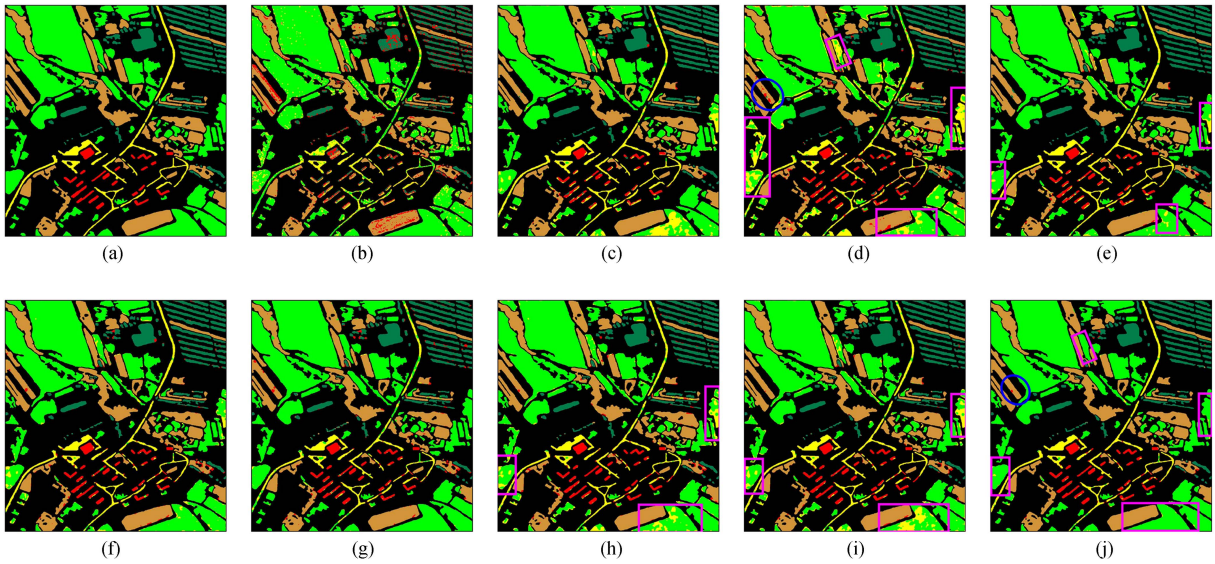

Fig. 13. Comparison classification maps on *Hebei_SL*. (a) Ground Truth. (b) WMM. (c) OMDF. (d) KP. (e) t-CNN. (f) FIL-Net. (g) CCR-Net. (h) GLT. (i) MFNet. (j) Ours.

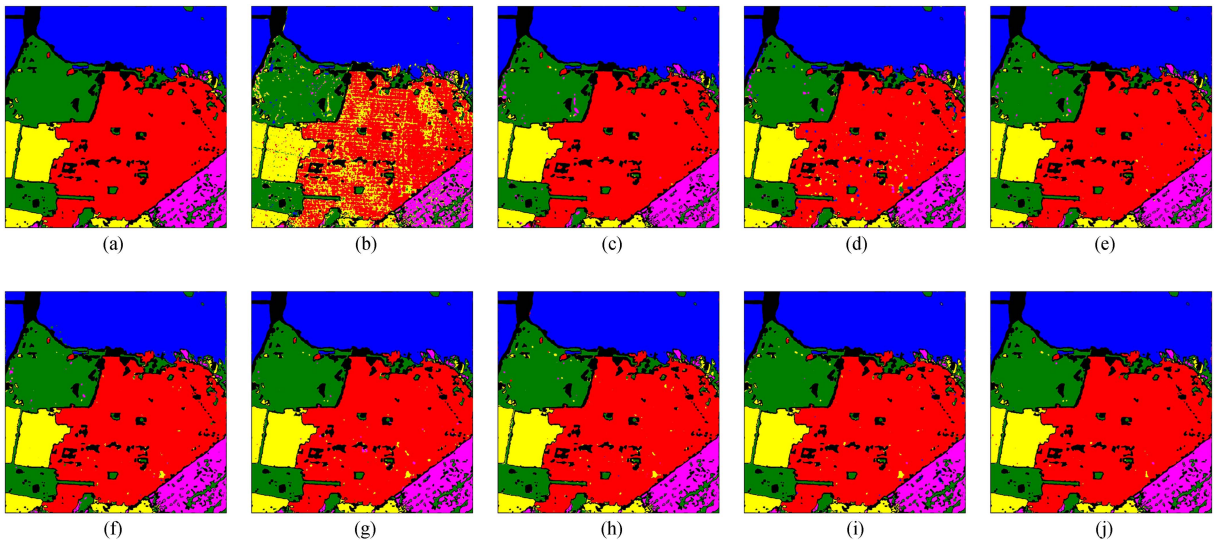

Fig. 14. Comparison classification maps on *SanFrancisco_CL*. (a) Ground Truth. (b) WMM. (c) OMDF. (d) KP. (e) t-CNN. (f) FIL-Net. (g) CCR-Net. (h) GLT. (i) MFNet. (j) Ours.

and MFNet have achieved high classification accuracy, but the proposed PolSAR-MPIformer can improve OA by 0.5%–0.7%, AA by 0.3%–0.7%, κ by 0.6%–0.8% on their basis, and achieves the highest accuracy in Wheat, Barley, and Fruits.

3) Results on Hebei_SL*:* The visual results on *Hebei_SL* are shown in Fig. 13. From Fig. 13, it can be seen that this dataset is prone to misclassify the Farmland in the lower right, middle left, and middle right regions. For example, in Fig. 13(c), (e), (h), and (i), some Farmland areas are misclassified as Road, all of them are highlighted with purple rectangles. It also can be seen from Fig. 8 that the feature differentiation between Road and Farmland is not strong enough. Besides, as can be seen from quantitative results in Table VIII that the t-CNN and proposed PolSAR-MPIformer have the highest classification accuracy in Road and Farmland, reaching 97.59% and 91.84% in Road, 96.55% and 99.05% in Farmland, respectively. However, t-CNN has low accuracy in Building, while PolSAR-MPIformer can reach 94.81% in Building. In addition, the proposed PolSAR-MPIformer achieves the highest accuracy in OA, AA, and κ. According to the above analysis, it is easy to conclude that the proposed PolSAR-MPIformer can obtain the best classification performance among these related algorithms on *Hebei_SL*.

4) Results on SanFrancisco_CL*:* The visual results and the quantitative results of different classification methods are presented in Fig. 14 and Table IX, respectively. From Fig. 14, it can be seen that the noise in classification maps of each algorithm is not severe, except for the traditional WMM. In Table IX, the classification accuracy of this dataset is significantly higher than that of the previous datasets, which is consistent with the strong

TABLE VIII
COMPARISON CLASSIFICATION ACCURACY ON *HEBEI_SL*

Method	WMM	OMDF	KP	t-CNN	FIL-CNN	CCR-Net	GLT	MFNet	Ours
Forest	79.65	99.32	96.38	**99.73**	97.39	95.10	97.18	98.65	98.65
Farmland	91.96	92.61	80.41	96.55	96.09	99.51	92.64	91.30	**99.05**
Road	56.19	94.75	69.07	**97.59**	92.72	80.70	89.23	92.33	91.84
Building	31.51	73.37	83.07	84.65	93.47	94.15	**96.98**	95.60	94.81
Bare land	91.65	97.23	93.48	94.24	97.14	96.39	96.13	97.70	**99.09**
OA(%)	86.51	94.23	85.18	94.07	96.25	96.87	94.02	94.06	**98.45**
AA(%)	70.19	91.45	84.48	93.57	95.36	93.17	94.43	95.12	**96.69**
$\kappa\times100$	79.09	91.14	78.12	90.89	94.17	95.04	90.84	90.94	**97.57**

The bold values represent the best results.

TABLE IX
COMPARISON CLASSIFICATION ACCURACY ON *SANFRANCISCO_CL*

Method	WMM	OMDF	KP	t-CNN	FIL-CNN	CCR-Net	GLT	MFNet	Ours
Forest	91.93	95.04	95.84	96.25	98.09	**98.57**	97.14	98.14	98.55
Water	99.30	99.90	99.69	99.95	98.88	**99.98**	99.84	99.83	99.86
High-density urban	75.94	**99.85**	97.06	99.57	99.22	98.74	99.26	99.28	99.73
Low-density urban	94.17	98.28	98.46	98.10	99.07	98.67	99.28	**99.33**	98.64
Developed	81.47	**98.96**	97.40	99.45	97.59	97.60	98.38	93.33	97.46
OA(%)	87.19	98.79	97.70	98.93	98.80	98.97	98.98	98.85	**99.31**
AA(%)	88.56	98.41	97.69	98.42	98.57	98.71	98.78	97.98	**98.85**
$\kappa\times100$	82.99	98.34	96.84	98.52	98.35	98.58	98.59	98.41	**99.05**

The bold values represent the best results.

compactness intraclass and strong separation interclass shown in Fig. 9. For DL-based classification algorithms, the accuracy of this dataset can reach over 98%. Despite the proposed PolSAR-MPIformer cannot achieve the highest accuracy in each category, it is only 0.02%–1.50% different from the highest accuracy. As for OA, AA, and κ, the proposed PolSAR-MPIformer achieves the highest value compared to other algorithms.

V. CONCLUSION

In this study, based on the overall framework of ViT, we propose a novel PolSAR-MPIformer including the MPI module and DAF module for dual-frequency PolSAR image classification. Among them, the MPI module replaces the high-complexity self-attention block in ViT with PI intra- and intersample. It can realize the extraction of global–local spatial information within samples and correlation information among samples, thus obtaining more discriminative feature representations under a low-complexity network structure. In addition, the DAF module is established as the classifier in PolSAR-MPIformer. It reduces the impact of speckle noise and utilizes the complementarity of dual-frequency data through the CLAF block and CFCF block, respectively. The ablation experiments on MPI module and DAF module have verified the effectiveness of MPI module in extracting advanced feature representations and DAF module in noise reduction and dual-frequency fusion. Besides, the comparative experiments with several state-of-the-art algorithms have shown that the proposed PolSAR-MPIformer could achieve impressive classification performance.

Considering the contradiction between the scarcity of labeled samples and the greed of DL for labeled samples restricts the application of DL algorithms in practical PolSAR image classification tasks, our future work will further investigate the use of dual-frequency PolSAR data for semisupervised or self-supervised learning, so as to improve the PolSAR image classification performance with few labeled samples. In addition, the unique statistical characteristic of PolSAR data is still not explored in DL, so we would like to investigate how to combine the statistical characteristics with DL model to improve the PolSAR image classification.

ACKNOWLEDGMENT

The authors would like to thank the Aerospace Information Research Institute, Chinese Academy of Sciences, for providing the *Hebei_SL* dataset. In addition, the authors would like to thank all the anonymous reviewers for their insightful and invaluable comments, which are helpful for our discussion.

REFERENCES

[1] A. Freeman and S. L. Durden, "A three-component scattering model for polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 3, pp. 963–973, May 1998.
[2] Y. Yamaguchi, T. Moriyama, M. Ishido, and H. Yamada, "Four-component scattering model for polarimetric SAR image decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 8, pp. 1699–1706, Aug. 2005.

[3] S. R. 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.

[4] H. Bi, J. Sun, and Z. Xu, "A graph-based semisupervised deep learning model for PoLSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2116–2132, Apr. 2019.

[5] H. Bi, F. Xu, Z. Wei, Y. Xue, and Z. Xu, "An active deep learning approach for minimally supervised PoLSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9378–9395, Nov. 2019.

[6] H. Bi, J. Yao, Z. Wei, D. Hong, and J. Chanussot, "PoLSAR image classification based on robust low-rank feature extraction and Markov random field," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4005205.

[7] R. Wang, Y. Nie, and J. Geng, "Multiscale superpixel-guided weighted graph convolutional network for polarimetric SAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3727–3741, 2024.

[8] Y. Jiang, M. Li, P. Zhang, X. Tan, and W. Song, "Unsupervised complex-valued sparse feature learning for PoLSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5230516.

[9] J. S. Lee and E. Pottier, *Polarimetric Radar Imaging : From Basics to Applications*. Boca Raton, FL, USA: CRC Press, 2009.

[10] H. Bi, J. Sun, and Z. Xu, "Unsupervised PoLSAR image classification using discriminative clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3531–3544, Jun. 2017.

[11] W. Song, M. Li, P. Zhang, Y. Wu, X. Tan, and L. An, "Mixture wg γ -mrf model for PoLSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 905–920, Feb. 2018.

[12] C. Liu, H.-C. Li, W. Liao, W. Philips, and W. Emery, "Variational textured dirichlet process mixture model with pairwise constraint for unsupervised classification of polarimetric SAR images," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4145–4160, Aug. 2019.

[13] W. An, Y. Cui, and J. Yang, "Three-component model-based decomposition for polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 6, pp. 2732–2739, Jun. 2010.

[14] Y. Cui, Y. Yamaguchi, J. Yang, H. Kobayashi, S.-E. Park, and G. Singh, "On complete model-based decomposition of polarimetric SAR coherency matrix data," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 1991–2001, Apr. 2014.

[15] 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.

[16] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, Dec. 2016.

[17] S.-W. Chen and C.-S. Tao, "Polsar image classification using polarimetric-feature-driven deep convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 627–631, Apr. 2018.

[18] X. Tan, M. Li, P. Zhang, Y. Wu, and W. Song, "Deep triplet complex-valued network for PoLSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10179–10196, Dec. 2021.

[19] Y. Jiang, M. Li, P. Zhang, and W. Song, "Semisupervised complex network with spatial statistics fusion for PoLSAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9749–9761, 2023.

[20] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–22.

[21] H. Dong, L. Zhang, and B. Zou, "Exploring vision transformers for polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5219715.

[22] 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, 2022, Art. no. 4505405.

[23] F. Fan et al., "Efficient instance segmentation paradigm for interpreting SAR and optical images," *Remote Sens.*, vol. 14, no. 3, Jan. 2022, Art. no. 531.

[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[25] H. Wang, C. Xing, J. Yin, and J. Yang, "Land cover classification for polarimetric SAR images based on vision transformer," *Remote Sens.*, vol. 14, no. 18, Sep. 2022, Art. no. 4656.

[26] Z. Pan, B. Zhuang, H. He, J. Liu, and J. Cai, "Less is more: Pay less attention in vision transformers," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 2035–2043.

[27] Z. Pan, J. Cai, and B. Zhuang, "Fast vision transformers with hilo attention," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 14541–14554, 2022.

[28] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10819–10829.

[29] 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, 2023, Art. no. 4006405.

[30] Y. Cao, Y. Wu, M. Li, W. Liang, and X. Hu, "Dfaf-Net: A dual-frequency PoLSAR image classification network based on frequency-aware attention and adaptive feature fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224318.

[31] 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, 2023, Art. no. 5205715.

[32] X. Xin et al., "Semi-supervised classification of dual-frequency PoLSAR image using joint feature learning and cross label-information network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5235716.

[33] X. Dupuis, V. Wasik, A. Alakian, and D. Dubucq, "Multi-band supervised classification for polarimetric SAR," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5772–5775.

[34] L. Ferro-Famil, E. Pottier, and L. Jong-Sen, "Unsupervised classification of multifrequency and fully polarimetric SAR images based on the h/a/alpha-wishart classifier," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 11, pp. 2332–2342, Nov. 2001.

[35] W. Gao, J. Yang, and W. Ma, "Land cover classification for polarimetric SAR images based on mixture models," *Remote Sens.*, vol. 6, no. 5, pp. 3770–3790, 2014.

[36] C. Liu, J. Yin, J. Yang, and W. Gao, "Classification of multi-frequency polarimetric SAR images based on multi-linear subspace learning of tensor objects," *Remote Sens.*, vol. 7, no. 7, pp. 9253–9268, Jul. 2015.

[37] F. Yang, W. Gao, B. Xu, and J. Yang, "Multi-frequency polarimetric SAR classification based on Riemannian manifold and simultaneous sparse representation," *Remote Sens.*, vol. 7, no. 7, pp. 8469–8488, Jul. 2015.

[38] S. De, D. Ratha, D. Ratha, A. Bhattacharya, and S. Chaudhuri, "Tensorization of multifrequency PoLSAR data for classification using an autoencoder network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 542–546, Apr. 2018.

[39] D. Hong et al., "Spectralgpt: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, Apr. 2024.

[40] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[41] X. He, Y. Chen, L. Huang, D. Hong, and Q. Du, "Foundation model-based multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502117.

[42] K. S. Chen, W. P. Huang, D. H. Tsay, and F. Amar, "Classification of multifrequency polarimetric SAR imagery using a dynamic learning neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 3, pp. 814–820, May 1996.

[43] T. Gadhiya and A. K. Roy, "Optimized wishart network for an efficient classification of multifrequency PoLSAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1720–1724, Nov. 2018.

[44] M. Ahishali, S. Kiranyaz, T. Ince, and M. Gabbouj, "Multifrequency PoLSAR image classification using dual-band 1D convolutional neural networks," in *Proc. Mediterranean Middle-East Geosci. Remote Sens. Symp.*, 2020, pp. 73–76.

[45] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[46] G. Liu, M. Li, Y. Wang, P. Zhang, Y. Wu, and H. Liu, "Four-component scattering power decomposition of remainder coherency matrices constrained for nonnegative eigenvalues," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 494–498, Feb. 2014.

[47] X. Nie, R. Gao, R. Wang, and D. Xiang, "Online multiview deep forest for remote sensing image classification via data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1456–1460, Aug. 2021.

[48] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.
[49] Z. Han, Y. Gao, X. Jiang, J. Wang, and W. Li, "Multisource remote sensing classification for coastal wetland using feature intersecting learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6008405.
[50] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.
[51] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Globallocal transformer network for HSI and LiDAR data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541213.
[52] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5404418.

Xinyue Xin received the B.S. degree in electrical information science and technology from Shaanxi Normal University, Xi'an, China, in 2019. She is currently working toward the Ph.D. degree in signal and information processing with the National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an.

Her main research interests are polarimetric synthetic aperture radar image analysis and interpretation and deep learning.

Ming Li (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in signal processing from Xidian University, Xi'an, China, in 1987, 1990, and 2007, respectively.

In 1987, he joined the Department of Electronic Engineering, Xidian University, where he is currently a Professor with the National Key Laboratory of Radar Signal Processing. His research interests include adaptive signal processing, detection theory, ultrawideband, and synthetic aperture radar image processing.

Yan Wu (Member, IEEE) received the B.S. degree in information processing and the M.S. and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1987, 1998, and 2003, respectively.

From 2003 to 2005, she was a Postdoctoral Fellow with the National Key Laboratory of Radar Signal Processing, Xi'an. Since 2005, she has been a Professor with the Department of Electronic Engineering, Xidian University. Her research interests include remote sensing image analysis and interpretation, data fusion of multisensor images, synthetic aperture radar autotarget recognition, and statistical learning theory and application.

Xiang Li received the B.S. degree in electronic and information engineering and M.S. degree in signal and information processing from Xidian University, Xi'an, China, in 2009 and 2012, respectively. He is currently working the Ph.D. degree with the Department of Electronic Engineering, Tsinghua University, Beijing, China.

He is a Researcher with the Beijing Institute of Radio Measurement. His research interests include radar system design, cognitive SAR/ISAR imaging, and moving target parameter estimation.

Peng Zhang (Member, IEEE) received the B.S. degree in electronic and information engineering, the M.S. and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 2006, 2009, and 2012, respectively.

He is currently an Associate Professor with National Key Laboratory of Radar Signal Processing, Xidian University. His main research interests are SAR image interpretation and statistical learning theory.

Dazhi Xu received the B.S. degree in electrical and information engineering from Lanzhou University of Technology, Lanzhou, China, in 2019. He is currently working toward the Ph.D. degree in signal and information processing with the National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, China.

His main research interests are polarimetric synthetic aperture radar image analysis and interpretation and deep learning.