Multifrequency PolSAR Image Fusion Classification Based on Semantic Interactive Information and Topological Structure

Yice Cao, Yan Wu*, Member, IEEE, Ming Li*, Member, IEEE, Mingjie Zheng, Peng Zhang*, and Jili Wang

Abstract—Compared with the rapid development of single-frequency polarimetric synthetic aperture radar (PolSAR) image classification technology, there is less research on the land cover classification of multifrequency PolSAR (MF-PolSAR) images. Also, the deep learning methods among them are mainly based on convolutional neural networks (CNNs), and only local spatiality is considered, but the nonlocal relationship is ignored. Therefore, this article proposes the multifrequency semantics and topological fusion (MF-STF) model based on semantic interaction and nonlocal topological structure to improve the MF-PolSAR classification performance. During MF-STF optimization, the semantic information-based classification (SIC) and topological property-based classification (TPC) work collaboratively, not only fully leveraging the complementarity of bands but also combining local and nonlocal spatial information to improve the discrimination of different categories. For SIC, the designed cross-band interactive feature extraction (CIFE) module is embedded to explicitly model the deep semantic correlation among bands, thereby leveraging the complementarity of bands to make ground objects more separable. In TPC, the graph sample and aggregate network (GraphSAGE) is employed to dynamically capture the representation of nonlocal topological relations between land cover categories. In this way, the robustness of classification can be further improved by combining nonlocal spatial information. Finally, a multifrequency weighted fusion (MWF) strategy is proposed to merge inference from different bands, so as to make the multifrequency (MF) joint classification decisions of SIC and TPC. Notably, its weights are adjusted based on the total model loss. The effectiveness of the proposed modules is proven by ablation experiments on three measured MF-PolSAR datasets. In addition, the comparative experiments show that MF-STF can achieve more competitive classification performance than some state-of-the-art methods.

Index Terms—Cross-band semantic interactive information, deep learning (DL), multifrequency polarimetric synthetic aperture radar (MF-PolSAR) image classification, multifrequency weighted fusion (MWF), topological structure.

Manuscript received 26 August 2022; revised 2 December 2022 and 14 February 2023; accepted 27 March 2023. Date of publication 5 April 2023; date of current version 14 April 2023. This work was supported in part by the National Science Foundation of China under Grant 62172321 and in part by the Civil Space Thirteen Five Years Pre-Research Project under Grant D040114. (Corresponding author: Yan Wu.)

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

Ming Li and Peng Zhang are with the National Key Laboratory of Radar Signal Processing, Xidian University, Xi’an 710071, China.

Mingjie Zheng and Jili Wang are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100045, China.

Digital Object Identifier 10.1109/TGRS.2023.3264560
so on. Based on the imaging difference analysis of PolSAR in different bands, combining multiple bands contains more abundant and comprehensive ground object information, which is theoretically beneficial to PolSAR image interpretation [7]. In addition, utilizing the information complementarity of bands is expected to eliminate the inaccuracy of object recognition in a single band.

The purpose of MF-PolSAR land cover classification is to determine the ground object coverage according to the characteristics of band data and the complementarity of bands [5]. Accurate MF-PolSAR image classification contributes to further promoting the application of PolSAR in civil and military fields. At present, the traditional methods for MF-PolSAR image classification, including statistical modeling [4] and [8], support vector machine (SVM) [24], Stein-sparse representation [7], and tensor representation [11], try to merge MF information to permit better classification performance than SF cases. However, their classification accuracy is limited due to the weak discriminative ability of handcrafted features [2]. Recently, DL-based methods have emerged to flexibly learn discriminative feature representation. For example, in [25], features extracted from the Kronecker product matrix were used as the input of an artificial neural network (ANN) for MF-PolSAR classification. In addition, in [2], by concatenating all bands as one band, the MF-PolSAR classification was carried out through the dilated convolution and pixel-refining parallel mapping network (CRPM-Net).

However, compared with the long research focus on SF-PolSAR image classification, the land cover classification of MF-PolSAR images has not been widely and deeply studied due to its particularity and challenges. There are two main reasons. On the one hand, the complementary information of bands has not been fully exploited. The features extracted by the existing MF-PolSAR classification methods [24], [25] are not enough to capture and reflect the complementarity, making it difficult to distinguish similar objects in MF classification results and failing to eliminate the classification inaccuracy under single bands. On the other hand, the MF information fusion methods with good generalization ability remain to be proposed. There are attribute differences among MF features, and simple fusion strategies adopted by existing methods [2], [7], [24], [25] may not be able to effectively resist and eliminate such differences. Therefore, robust classification results may not be obtained. In addition, the research on MF information fusion strategies that are not specific to processing data and the number of bands is the necessary technical support for the practical application of future MF-PolSAR systems. Therefore, how to leverage the complementarity and effectively fuse MF information is the key to achieve more satisfactory classification results under MF conditions than under SF conditions.

In addition, notably, the existing neural networks [2], [25] for MF-PolSAR image classification are all based on the CNN framework. However, CNN performs convolution on regular regions with a fixed-size convolution kernel, so it can only model short-range local spatial relationships [26], which will limit the improvement of classification performance. To address this problem, based on the graph structure to model the topological relationship between samples, graph convolutional networks (GCNs) [27] have been proposed and have shown great advantages in many applications, including PolSAR image classification [30], [31], [32]. The topological relationship describes the overall characteristics between samples (nodes), which belong to the medium- and long-range context relationship and are not limited by the coordinates in images [28]. Therefore, the use of topological structure can effectively capture the nonlocal spatial information, which better describes the overall characteristics between categories and, in turn, helps to improve the classification performance [29]. Worthy, the spatial-based GCNs [33], [34], [35] have attracted much attention due to their better flexibility and efficiency. This kind of GCN defines graph convolution based on the spatial relationship of nodes, which can flexibly aggregate new nodes and reduce computational complexity. Among them, the graph sample and aggregate network (GraphSAGE) [34] with inductive learning ability has good generalization. It makes the distributed training of large-scale graph data possible and expands the application scope of GCN. So far, spatial-based GCNs have not been used in MF-PolSAR image classification.

Based on the above analysis, to further enhance the MF-PolSAR classification performance, a novel MF-PolSAR classification model named MF semantics and topology fusion (MF-STF) is proposed in this article. The proposed MF-STF is mainly based on semantic interactive information and topological structure and can fully mine and leverage the complementarity of bands as well as combine local and nonlocal spatial information to improve the accuracy and robustness of classification. During MF-STF optimization, the semantic information-based classification (SIC) and the topological property-based classification (TPC) are adopted for each band, which work cooperatively for more accurate model learning. To fully mine and leverage the complementarity of bands, a shared part named cross-band interactive feature extraction (CIFE) module is adopted and embedded in SIC. CIFE is designed to explicitly model the deep semantic correlation among bands by using the correlations between attributes of band-specific representations. It can extract more discriminative cross-band interactive features and enhance the interactive fusion of information between bands, thereby improving the accuracy of classifying ground objects.

In addition, TPC mainly utilizes two layers of GraphSAGE to dynamically capture the representation of nonlocal topological relations between samples, thus further improving the robustness of classification by combining nonlocal spatial information. Finally, the MF weighted fusion (MFWF) strategy is adopted to merge inference from different bands in an adaptive weighted manner, thereby making the MF joint classification decisions for SIC and TPC. The weights of MFWF related to different bands are adaptively updated based on the total model loss. It should be emphasized that a consistency loss is added to the total loss to make SIC and TPC collaboratively work better during the model optimization process and to make their prediction results closer. In addition, to achieve efficient MF-PolSAR image classification prediction, only the SIC results
of each band are combined as the final MF joint classification result.

The crucial contributions can be summarized as follows.

1) An MF-PolSAR image classification model named MF-STF is proposed, including two kinds of classifications, SIC and TPC. SIC and TPC work collaboratively during model optimization, which can not only make full use of the complementary and interaction information among bands but also combine local and nonlocal spatial information, thereby obtaining more accurate and robust classification results.

2) The designed CIFE module is used in SIC to explicitly model semantic interaction between different bands. It is an important part of MF feature learning and the key to improving classification performance. The interactive features extracted by CIFE are concatenated with band-specific features to realize the SIC, which improves the discrimination between different categories, thereby leading to better classification performance.

3) The GraphSAGE is adopted in TPC to dynamically capture the representation of nonlocal topological relations, so as to further improve the robustness of classification.

4) The MFWF strategy with good generalization is proposed to flexibly fuse the results of different bands, thereby adaptively obtaining the final MF joint classification decision. Notably, its weights are updated adaptively based on the total model loss.

The rest of this article is organized as follows. Section II describes the proposed MF-STF model for MF-PolSAR image classification. In Section III, experimental results and analysis on three measured MF-PolSAR datasets are presented. The conclusion is drawn in Section IV.

II. PROPOSED MF-STF MODEL FOR MF-POLSAR IMAGE CLASSIFICATION

For brevity and clarity, taking the case of two frequency bands (A-band and B-band) as an example, Fig. 1 intuitively shows the architecture of the proposed MF-STF model for MF-PolSAR image classification. On this basis, the architecture of more than two bands can be extended. As shown in Fig. 1, the proposed model mainly concludes band-specific semantic feature extraction (BSFE), CIFE, SIC, TPC, and the final MF joint classification decision based on the MFWF strategy. In addition, Table I reports the detailed structure of MF-STF.

### Table I

| Structure | Output Size |
|-----------|-------------|
| BSFE      |             |
| A-band    | 9 x 13 x 13 |
| B-band    | 9 x 13 x 13 |
| CIFE      |             |
| A-band    | 32 x 13 x 13|
| B-band    | 32 x 13 x 13|
| Concatenation | 96 x 13 x 13 |
| SIC       |             |
| A-band    | Global Pool / FC / Softmax |
| B-band    | Global Pool / FC / Softmax |
| TPC       |             |
| A-band    | GraphSAGE(64-32) / FC / Softmax |
| B-band    | GraphSAGE(64-32) / FC / Softmax |
| Multi-frequency Joint Prediction | MFWF / Softmax |

A. Band-Specific Semantic Feature Extraction

The BSFE part aims to capture various band-specific information for MF learning enhancement, thereby preserving the interband discriminative representation and ensuring the diversity of band attributes.

In the monostatic backscattering case, the scattering characteristics of each resolution cell in a PolSAR image can be described by the $3 \times 3$ polarimetric coherency matrix $T$ [1]

$$T = \langle u_L \cdot u_L^H \rangle = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}$$ (1)

where $\langle \cdot \rangle$ and $H$ denote ensemble averaging and conjugate transpose, respectively, and $u_L$ is the polarimetric target vector. In this article, a 9-D vector expended by the upper triangular...
of \( T \) is used as the input feature of each pixel cell and can be represented by
\[
[T_{11}, T_{22}, T_{33}, \Re(T_{12}), \Re(T_{13}), \Re(T_{23}), \Im(T_{12}), \Im(T_{13}), \Im(T_{23})]
\]
where \( \Re(.) \) and \( \Im(.) \) denote the real and imaginary parts of complex elements, respectively.

Suppose that there are \( K \) bands. First, \( K \) small data patches of size \( 9 \times n \times n \) centered on the same pixel in all band images are generated as input to the BSFE part. \( n \times n \) is the spatial dimension of small data patch, which is set to \( 13 \times 13 \) in this article. For each band, the band-specific CNN in BSFE consists of three cascaded convolution blocks, and its corresponding structure is shown in Fig. 1 and Table I. Let \( X_k \in \mathbb{R}^{9 \times n \times n} \) be an input data patch of the \( k \)th band, and the output features \( X_{k}^{b1} \) obtained by the first convolution block \( F_{b1}(\cdot) \) can be formulated as
\[
X_{k}^{b1} = F_{b1}(X_k) = f_{b1}^{\text{ReLU}} \left(f_{b1}^{\text{BN}} \left(f_{b1}^{\text{Conv}}(X_k)\right)\right)
\]
where \( k = \{1, \ldots, K\} \), \( f_{b1}^{\text{Conv}}(\cdot), f_{b1}^{\text{BN}}(\cdot), \) and \( f_{b1}^{\text{ReLU}}(\cdot) \) denote the convolution, batch normalization (BN), and rectified linear unit (ReLU) in the first convolution block of CNN, respectively. The other two convolution blocks \( F_{b2}(\cdot) \) and \( F_{b3}(\cdot) \) have the same processing as \( F_{b1}(\cdot) \). The only difference between these three blocks lies in the output dimension. Therefore, the corresponding BSFE output \( X_{k}^{\text{BSFE}} \in \mathbb{R}^{64 \times n \times n} \) of the \( k \)th band can be formulated as
\[
X_{k}^{\text{BSFE}} = F_{b3}(F_{b2}(F_{b1}(X_k))).
\]

The output features \( \{X_{k}^{\text{BSFE}}\}_{k \in \{1, \ldots, K\}} \) extracted by the BSFE part can not only preserve the discriminative representation between bands but also ensure the interband diversity, which is a key aspect of MF learning enhancement.

**B. Cross-Band Interactive Feature Extraction**

The features extracted by BSFE are only specific to bands, and there is no information interaction between bands. This may not fully utilize the complementary information of classifications, thereby limiting the improvement of classification performance. To focus on the correlation of MF-PolSAR data, based on the band-specific semantic representation of BSFE, the CIFE module is utilized to explicitly model the deep semantic correlation between bands, thereby extracting discriminative enough features to improve category differentiation.

Formally, with respect to the \( k \)th band, a set \( S^k \) containing different band pairs is denoted as
\[
S^k = \{(k, \bar{k})\}_{\bar{k} \in \{1, \ldots, K\} \setminus k}
\]
where \( (k, \bar{k}) = (\bar{k}, k) \) means undirected. Based on the semantic representation extracted by BSFE, the correlative feature maps \( X_{k, \bar{k}} \) between \( X_{k}^{\text{BSFE}} \) and \( X_{\bar{k}}^{\text{BSFE}} \) can be formulated as
\[
X_{k, \bar{k}} = F_{\text{Cor}}(X_{k}^{\text{BSFE}}, X_{\bar{k}}^{\text{BSFE}})
\]
where \( F_{\text{Cor}}(\cdot) \) corresponds to the correlation operation. In detail, assume that \( X_{k}^{\text{BSFE}} \) represents any feature map in \( X_{k}^{\text{BSFE}} \), where \( i = \{1, \ldots, I\} \). \( I \) denotes the number of feature maps. We first do a dot product between \( X_{k}^{\text{BSFE}} \) and each feature map of \( X_{k}^{\text{BSFE}} \) to obtain a set of feature maps \( X_{k, \bar{k}} \), which can be formulated as
\[
X_{k, \bar{k}} = X_{k}^{\text{BSFE}} \ast X_{\bar{k}}^{\text{BSFE}}
\]
where \( \ast \) denotes the dot product operator. Thus, for all \( I \) feature maps in \( X_{k}^{\text{BSFE}} \), the correlative feature maps \( X_{k, \bar{k}} \) can be described as
\[
X_{k, \bar{k}} = \{x_{k, \bar{k}}^{i}\}_{i \in \{1, \ldots, I\}}.
\]

Therefore, based on the above process, for the \( k \)th band, all correlative feature maps with respect to other bands can form a set \( D^k_S = \{X_{k, \bar{k}}^{i}\}_{(k, \bar{k}) \in S^k} \).

Finally, to reduce computational complexity, a shared convolution block \( F_b(\cdot) \) with \( 1 \times 1 \) convolutional kernel and 32 output dimension is adopted to project each \( X_{k, \bar{k}} \) in \( D^k_S \). The projected output \( X_{k, \bar{k}}^{\text{Cor}} \) is
\[
X_{k, \bar{k}}^{\text{Cor}} = F_b(X_{k, \bar{k}}) = f_{b}^{\text{ReLU}} \left(f_{b}^{\text{BN}} \left(f_{b}^{\text{Conv}}(X_{k, \bar{k}})\right)\right)
\]
where \( f_{b}^{\text{Conv}}(\cdot), f_{b}^{\text{BN}}(\cdot), \) and \( f_{b}^{\text{ReLU}}(\cdot) \) denote the convolution, BN, and ReLU in \( F_b(\cdot) \), respectively.

Thus, for the \( k \)th band, the correlative feature maps with respect to other bands are collected in a set \( \{X_{k, \bar{k}}^{\text{Cor}}\}_{(k, \bar{k}) \in S^k} \).

We concatenate all elements in this set as the final cross-band interactive feature maps \( X_{k}^{\text{CIFE}} \) with respect to the \( k \)th band.

For more clarity, we illustrate the CIFE module of two bands, as shown in Fig. 2. Respecting to A-band, the correlative feature maps between A- and B-bands are calculated by the traversing multiplication between each feature map of A and all feature maps of B. In Fig. 2, this calculation process is highlighted by the blue dashed rectangle. For the B-band, the correlative feature maps are calculated by the traversing multiplication between each feature map of B and all feature maps of A, and the calculation process is highlighted by the red dashed rectangle in Fig. 2. Finally, these correlative features are projected by the shared convolution block to obtain the final cross-band interactive features. Notably, when the number of frequency bands is greater than 2, for any band, the final cross-band interactive features are the concatenation of the calculation results of the current band and all other bands.

The CIFE module can utilize the correlation between the attributes of semantic representation to explicitly model the
deep semantic interaction between bands. The more discriminative cross-band interactive features extracted by the CIFE module realize the interactive fusion and enhancement of information among bands, thereby improving the discrimination between different categories and leading to better classification performance.

C. Semantic Information-Based Classification

Based on the combination of band-specific and interactive information, SIC is designed to obtain the category prediction from the semantic and local spatial perspectives, which can fully utilize the complementarity of bands to improve the distinction of ground object categories.

For each band, SIC consists of a global average pooling (GAP) layer, a fully connected (FC) layer with C output, and a Softmax layer, where C is the total number of categories. Representing the concatenated features of kth band as $X^\text{Con}_k$, the output $Z^\text{SIC}_k$ of SIC can be denoted as

$$Z^\text{SIC}_k = f^\text{Softmax} (f^\text{FC} (f^\text{GAP} (X^\text{Con}_k)))$$  \hfill (10)

where $f^\text{GAP} (\cdot)$, $f^\text{FC} (\cdot)$, and $f^\text{Softmax} (\cdot)$ respectively the GAP, FC, and Softmax operations in the SIC part.

D. Topological Property-Based Classification

SIC based on CNN frameworks is unable to model the nonlocal topological relationship between land cover categories that describes the overall characteristics, which limits the further improvement of classification performance. To this end, TPC is designed to dynamically capture the representation of topological relations between samples, thereby further improving the classification results with the assistance of nonlocal spatial information [29].

Specifically, for each band, the two layers of GraphSAGE model are adopted in TPC. First, the nonlocal subgraph of node relationship for each band is constructed and initialized based on the training samples. Suppose that the concatenated semantic features of the sample $v$ in any band are denoted by $X^\text{Con}_v$, $u$ is the neighborhood nodes of $v$ in the constructed nonlocal graph, and $N(\cdot)$ represents the neighborhood function. After aggregating the information of neighborhood nodes in the first layer of GraphSAGE, the features of node $v$ are updated as

$$G^\text{TPC}_1(v) \leftarrow \sigma \left( W^\text{TPC}_1 \cdot F_{\text{Mean}} \left( X^\text{Con}_v \cup \left\{ X^\text{Con}_u : u \in N(v) \right\} \right) \right)$$  \hfill (11)

where $W^\text{TPC}_1$ represents the weights of the first layer and $\sigma$ and $F_{\text{Mean}}(\cdot)$ are, respectively, the nonlinear activation and mean aggregation functions. Then, after the information aggregation in the second layer of GraphSAGE, the updated features are

$$G^\text{TPC}_2(v) \leftarrow \sigma \left( W^\text{TPC}_2 \cdot F_{\text{Mean}} \left( G^\text{TPC}_1(v) \cup \left\{ G^\text{TPC}_1(u) : u \in N(v) \right\} \right) \right).$$  \hfill (12)

Finally, the features $G^\text{TPC}_2(v)$ will pass through an FC layer and a Softmax layer to obtain the final classification result of TPC

$$Z^\text{TPC}(v) = f^\text{Softmax} \left( f^\text{FC} \left( G^\text{TPC}_2(v) \right) \right).$$  \hfill (13)

Fig. 3 shows the information aggregation and propagation in simple two layers of GraphSAGE. For the kth band, based on the concatenated features $X^\text{Con}_v$ and the initial dynamic subgraph in Layer 0, GraphSAGE aggregates neighborhood nodes to update the feature representation of each central node. Through the repeated aggregation of two layers of GraphSAGE model, the node information can be extended to the second-order neighborhood. For example, in Fig. 3, the central node $a$ in Layer 2 can aggregate the information of nodes $e$ and $f$ in its second-order neighborhood ($L = 2$). Accordingly, the feature representation of nodes is aggregated and updated through information propagation to capture nonlocal spatial information, thereby improving the robustness of classification.

TPC works collaboratively with SIC during MF-STF optimization, so the proposed model can simultaneously capture the local and nonlocal spatial context information. In this way, more discriminative feature representation can be effectively obtained, which contributes to the accuracy and robustness of the classification.

E. Final MF Fusion Classification Decision Based on the MFWF Strategy

For each band, there are two probability outputs of SIC and TPC. To obtain the final MF joint classification decisions of SIC and TPC, the MFWF strategy is adopted to adaptively and flexibly combine the results of all bands.

For the kth band, two probability outputs $Z^\text{SIC}_k$ and $Z^\text{TPC}_k$ are obtained. Assume that $\alpha \in \mathbb{R}^K$ denotes the adaptive weight, and the MF joint classification results of SIC and TPC calculated by MFWF are, respectively, defined as

$$Y^\text{SIC} = \sum_{k=1}^{K} \alpha_k^r Z^\text{SIC}_k$$ \hfill (14)

$$Y^\text{TPC} = \sum_{k=1}^{K} \alpha_k^s Z^\text{TPC}_k$$ \hfill (15)

where $\sum_{k=1}^{K} \alpha_k = 1$, $\alpha_k \geq 0$. $\gamma > 1$ denotes the power exponent parameter of weights, which is used to avoid the trivial solution of $\alpha_k$ [36].
Then, the Softmax function $F_{\text{Softmax}}(\cdot)$ is utilized to calculate the probability distribution of each category, which can be described as

$$F_{\text{Softmax}}(Y_{\text{SIC}})_c = e^{y_{\text{SIC}}_c} / \sum_{j=1}^{C} e^{y_{\text{SIC}}_j}, \quad c = 1, \ldots, C$$  \hspace{1cm} (16)

$$F_{\text{Softmax}}(Y_{\text{TPC}})_c = e^{y_{\text{TPC}}_c} / \sum_{j=1}^{C} e^{y_{\text{TPC}}_j}, \quad c = 1, \ldots, C.$$  \hspace{1cm} (17)

### F. Parameter Updating

Two types of parameters need to be updated in the optimization process of the MF-STF model. One is the parameters of the BSFE, CIFE, SIC, and TPC modules, which can be called network parameters. The other is the weights $\alpha$ of MFWF.

1) Update Network Parameters: To update the network parameters, the cross-entropy loss [37] is first used in both SIC and TPC to calculate the semantic and topological classification losses, respectively. Based on the MFWF strategy, the two losses $\ell_{\text{SIC}}$ and $\ell_{\text{TPC}}$ are denoted as

$$\ell_{\text{SIC}} = \sum_{k=1}^{K} a_k L_k(Z_{\text{SIC}}^k, \text{Label})$$  \hspace{1cm} (18)

$$\ell_{\text{TPC}} = \sum_{k=1}^{K} a_k L_k(Z_{\text{TPC}}^k, \text{Label})$$  \hspace{1cm} (19)

where $L_k$ represents the cross-entropy loss function and Label denotes the true label of samples.

In addition, to make SIC and TPC collaboratively work better during model optimization and make their classification results closer, a consistency loss is used [38] to constrain the MF joint probability predictions of SIC and TPC, which is described as

$$\ell_{\text{consistency}} = \frac{1}{N} \|Y_{\text{SIC}} - Y_{\text{TPC}}\|_2$$  \hspace{1cm} (20)

where $N$ is the number of samples. For any sample, the smaller the $\ell_{\text{consistency}}$ value, the more consistent the prediction results of SIC and TPC, which reflects the better cooperation between SIC and TPC. Thus, the final objective function, that is, the total loss, is

$$\ell_{\text{total}} = \ell_{\text{SIC}} + \ell_{\text{TPC}} + \lambda \cdot \ell_{\text{consistency}}$$  \hspace{1cm} (21)

where $\lambda$ is the regularization parameters. In this article, $\lambda$ is set to 0.1. Based on the total loss, the network parameters can be updated by the backpropagation algorithm.

2) Update Weights $\alpha$: According to the total loss $\ell_{\text{total}}$, only $\ell_{\text{SIC}}$ and $\ell_{\text{TPC}}$ are related to $\alpha$. Therefore, by fixing the network parameters, the optimization problem for updating $\alpha$ is

$$\min_{\alpha} (\ell_{\text{SIC}} + \ell_{\text{TPC}}) = \min_{\alpha} \sum_{k=1}^{K} a_k L_k(Z_{\text{SIC}}^k, \text{label}) + L_k(Z_{\text{TPC}}^k, \text{label}).$$  \hspace{1cm} (22)

Here, we use $L_k$ to represent $L_k(Z_{\text{SIC}}^k, \text{label}) + L_k(Z_{\text{TPC}}^k, \text{label})$. According to $\sum_{k=1}^{K} a_k = 1$, $a_k \geq 0$, the corresponding Lagrangian function is

$$\omega(\alpha, \zeta) = \sum_{k=1}^{K} a_k^\gamma L_k - \zeta \left( \sum_{k=1}^{K} a_k^\gamma - 1 \right)$$  \hspace{1cm} (23)

where $\zeta$ is the Lagrange multiplier. According to (23), taking the derivatives with respect to $a_k$ and $\zeta$, the updating equation of weight $a_k$ can be obtained by setting derivatives to zero

$$a_k = \frac{L_k^{1/1-\gamma}}{\sum_{m=1}^{K} L_m^{1/1-\gamma}}.$$  \hspace{1cm} (24)

### G. MF-STF Optimization and Prediction

During MF-STF optimization, MF-PolSAR training samples are first input to BSFE to capture band-specific semantic features. Then, based on these semantic representations, CIFE is used to extract the cross-band interactive features of each band. For any band, two classification results are obtained from SIC and TPC based on the concatenation of band-specific semantic features and cross-band interactive features. After that, MFWF is adopted to combine the results of all bands for the final MF joint classification decisions of SIC and TPC. Finally, the model parameters are updated according to the calculated total loss. The above process is repeated continuously until the optimization termination condition is reached, and the entire MF-STF model optimization is completed.

In the prediction of the whole MF-PolSAR image, the number of prediction samples is much larger than the number of training samples. TPC requires a fixed number of samples to construct subgraphs of the same scale as in model optimization to achieve prediction, thus greatly increasing the time of classification prediction. In contrast, the prediction of SIC is not subject to this restriction and is simpler and more flexible than TPC. Therefore, the prediction time of SIC is shorter than that of TPC. Moreover, the MF joint prediction results of SIC and TPC are closer due to the constraint of consistency loss in model optimization. Therefore, to achieve efficient MF-PolSAR image classification, only the SIC results of each band are combined as the final MF joint classification prediction, that is, any pixel in the MF-PolSAR image can get the corresponding label prediction only through BSFE, CIFE, and SIC parts.

### III. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conduct experiments to evaluate the performance of the proposed MF-STF on three measured MF-PolSAR datasets.

#### A. Experimental Datasets Description

The three experimental MF-PolSAR datasets include the dual-band San Francisco, the dual-band Woniupan, and the three-band Flevoland datasets.

1) San Francisco: The San Francisco dataset is composed of the full-polarimetric L-band ALOS and C-band GaoFen-3 data. Its image size is $1161 \times 1161$. According to the corresponding optical image, the dataset contains five categories: forest, water, high-density urban, low-density urban, and developed. Fig. 4 shows the Pauli RGB images and the ground-truth image.
Fig. 4. Dual-frequency full PolSAR image over the San Francisco region. (a) C-band Pauli. (b) L-band Pauli. (c) Ground truth.

Fig. 5. Dual-frequency full PolSAR image over the Woniupan region. (a) S-band Pauli. (b) L-band Pauli. (c) Ground truth.

Fig. 6. MF full PolSAR image over the Flevoland region. (a) C-band Pauli. (b) L-band Pauli. (c) P-band Pauli. (d) Ground truth.

2) Woniupan: The Woniupan dataset was acquired by an airborne system in 2021, which concludes S- and L-bands. Its image size is $1005 \times 962$, and the Pauli RGB images are shown in Fig. 5(a) and (b). There are five land cover categories identified in this dataset: road, building, farmland, forest, and bareland. Fig. 5(c) shows the ground-truth image.

3) Flevoland: The Flevoland dataset is acquired by the NASA/JPL AIRSAR system in the C-, L-, and P-bands. It contains 15 categories and has a size of $1079 \times 1024$. The Pauli RGB and the ground-truth images are shown in Fig. 6.

B. Experimental Setup and Evaluation Criteria

For the nonoverlapping collection of the training and test samples, we adopt the chessboard-like sampling strategy [39] rather than the random sampling strategy. The chessboard-like sampling strategy can greatly reduce the spatial correlation between training and test data, thereby making the classification results not overly optimistic and making algorithms more suitable for practical applications. Its key step is the nonoverlapping data partition, whose simple demonstration is shown in Fig. 7. In Fig. 7, the large image is segmented into nonoverlapping chessboard-like subimages. Among them, the subimages corresponding to all-black blocks are one part, and the subimages corresponding to all-white blocks are the other part. The two nonoverlapping parts are alternately used as training data and test data. The results obtained when each part is used as test data are stitched as the final classification of the entire image. Specifically, for each experimental image, 400 nonoverlapping small images of the same size are divided. We first take 200 of them as training images and another 200 for the test. Then, do the same after swapping. We randomly choose 200 labeled pixels in training images to learn networks and stitch the two test results as the final classification.

In addition, to avoid overfitting, five data augmentation strategies are utilized to expand the training dataset, including horizontal flip, vertical flip, random 90°, 180°, and 270° rotations. The center pixels with $13 \times 13$ neighborhoods are used as the input patches. Adam with momentum 0.9 is used to update network parameters. The training epoch is set to 150 for all datasets. The learning rate and the batch size are set to 0.001 and 100, respectively. The proposed MF-STF is implemented on the PyTorch framework. All the experiments were run on a Lenovo Y720 cube gaming PC with an Intel Core i7-7700 CPU, an Nvidia GeForce GTX 1080 GPU, and 16-GB RAM under Ubuntu 20.04 LTS operating system.

To quantify the classification performance, the accuracy of each category, overall accuracy (OA), average accuracy (AA), and kappa coefficient ($\kappa$) are employed as evaluation metrics. In addition, to ensure fairness and reliability, all experiments are conducted ten times, and the mean values of evaluation metrics are reported.

C. Performance Analysis of the Proposed MF-STF

1) Effect of the Regularization Parameter $\lambda$: The regularization parameter $\lambda$ is used to adjust the weight of the consistency loss in the total loss. We vary $\lambda$ to 0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, and 1 to explore its influence on classification results. Fig. 8 shows the relationship between OA and different $\lambda$'s. Here, the power exponent parameter $\gamma$ is fixed to 3 for all experimental datasets.

As shown in Fig. 8, for the San Francisco and Flevoland datasets, the value of $\lambda$ has few effects on the classification performance. However, for the Woniupan dataset, when the value of $\lambda$ is too large, the classification performance will...
TABLE II

| OA (%) of the Proposed MF-STF With Different γ Values |
|---|---|---|---|---|---|---|---|---|---|
| γ  | 1.5 | 2   | 2.5 | 3   | 3.5 | 4   | 4.5 | 5   | 6   | 7   | 8   |
| SanFrancisco | 97.77 | 98.26 | 98.38 | 98.62 | 98.76 | 98.51 | 98.68 | 98.09 | 98.69 | 98.46 | 98.66 | 85.36 | 76.39 | 44.01 |
| Woniupan      | 91.16 | 92.82 | 93.93 | 94.41 | 93.92 | 95.41 | 94.62 | 92.44 | 91.46 | 89.75 | 87.70 | 70.02 | 64.84 | 61.64 |
| Flevoland     | 94.65 | 95.44 | 95.39 | 95.45 | 95.41 | 93.97 | 56.01 | 43.92 | 21.03 | 8.76  | 8.02  | 3.30  | 7.98  | 3.61  |

2) Effect of the Power Exponent Parameter γ: The power exponent parameter γ is one of the key parameters affecting the proposed MF-STF performance. As mentioned earlier, the value of γ must be greater than 1. Therefore, for the three experimental datasets, the optimal γ is selected from {1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8}. Table II shows the classification accuracy under the varying γ on these datasets. Notably, here, the regularization parameter λ is fixed to 0.1 for all experimental datasets.

As shown in Table II, for the SanFrancisco dataset, the results show that MF-STF achieves the highest results when γ is 3.5. For the Woniupan dataset, the best classification index is obtained when γ is 4. In addition, for the Flevoland dataset, the best result is realized when γ is 3. Therefore, γ = 3.5, γ = 4, and γ = 3 are the best choice for the three datasets.

3) Ablation Experiment: As analyzed earlier, the proposed MF-STF mainly relies on CIFE and TPC to enhance the classification performance. Therefore, we conduct the related ablation experiment to verify the effectiveness of the two key components. Table III shows the ablation results, where √ and ×, respectively, denote with and without this part.

As shown in Table III, it is obvious that the complete MF-STF model with all key parts achieves the highest results. Without CIFE and TPC, Model1 achieves the lowest classification results on all datasets. When considering CIFE, for the three datasets, Model2 outperforms Model1 by 0.79%, 2.95%, and 1.51% on OA. This result safely demonstrates the superiority of capturing interactive information between bands. In addition, compared with Model1, Model3 with TPC achieves 1.14%, 2.4%, and 0.48% improvements on OA for the three datasets, which demonstrates the effectiveness of nonlocal topological information in improving classification performance. It should be noted that MF-STF outperforms Model2 and Model3. This suggests that the simultaneous extraction of interactive and topological information can provide more comprehensive information for classification interpretation.

To sum up, the experimental results shown in Table III demonstrate the effectiveness of CIFE and TPC. In addition, they also demonstrate that the proposed MF-STF can obtain a more accurate classification due to simultaneously considering interactive and topological information.

4) Validation of CIFE: To further verify the effectiveness of the CIFE module, we construct a comparison module composed of two common convolution blocks, whose input is the concatenated features of different bands output by the BSFE part. The convolution output dimensions are set to 64 and 32. The comparison model using this module is called Model_common. Table IV shows the classification results of Model_common and Model2. In addition, Fig. 9 shows the t-SNE visualization of features extracted by Model_common and Model2. Notably, these features are the MF joint output features of networks in the prediction stage.

As shown in Table IV, Model2 outperforms Model_common on all datasets. Compared with Model_common, Model2 could increase OA by 0.3%–1.98%, AA by 0.59%–2.65%, and Kappa by 0.004–0.0234. In addition, as shown in Fig. 9, the feature distribution between categories in Net2 is more

Fig. 8. Classification accuracy of the proposed MF-STF with different values of the regularization parameter λ on the three experimental datasets.

decline significantly. Therefore, to ensure the generalization, the regularization parameter λ for the three experimental datasets is all selected as 0.1.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
separable than in Model_common. Specifically, for the San-
Francisco dataset, the feature distribution clusters of Cate-
gories 1 and 2 in Fig. 9(c) are more compact than those in
Fig. 9(b). For the Woniupan dataset, Categories 4 and 5 in
Fig. 9(h) can be better distinguished from other categories
than in Fig. 9(g). For the Flevoland dataset, compared with
Fig. 9(l), the distribution overlap of Category 15 with other
categories in Fig. 9(m) is reduced. Furthermore, the feature
distribution within categories in Fig. 9(m) is more compact
than in Fig. 9(l). Therefore, summarizing the above analysis,
we can conclude that the proposed CIFE has advantages over
common convolution and is more effective for improving the
classification performance.

5) Validation of TPC: Fig. 9 also visualizes the feature
extracted by Model1, Model3, and the proposed MF-STF to
validate the importance of TPC.

As shown in Fig. 9, for the three datasets, the feature
discrimination between categories in Model3 is improved
than in Model1. Comparing Fig. 9(a) with Fig. (d), it can be
seen that the compactness within categories in Fig. 9(d) is
better than in Fig. 9(a). For the Woniupan dataset, compared
with Fig. 9(i), there are clear distribution boundaries between
categories in Fig. 9(i), and Categories 2 and 4 in Fig. 9(i)
can be better distinguished from other categories. In addition,
the distances between category distribution clusters in
Fig. 9(n) are relatively farther than in Fig. 9(k). These results
illustrate that the features extracted by Model3 are more
discriminative than those by Model1, which can prove that
the use of TPC is effective for better category distinction
and improvement of classification performance. Notably, for
each dataset, the feature distribution obtained by MF-STF
is better than other networks, which indicates that the network
structure we constructed is reasonable and effective, and the
proposed modules make the learned features of MF-STF more
discriminative.

6) Validation of the Consistency Loss: We compare the
prediction results and prediction times of SIC and TPC with
or without the consistency loss. Table V summarizes the
experimental results on the three datasets.

As shown in Table V, the OA of SIC and TPC with consis-
tency loss is higher than without consistency loss. In addition,
the prediction results of SIC and TPC are close, especially
when the consistency loss is employed. Specifically, for the
SanFrancisco dataset, the OA differences between SIC and
TPC with and without consistency loss are 0.1% and 0.75%,
respectively. For the Woniupan dataset, such OA differences
are 0.4% and 0.75%. For the Flevoland dataset, such OA
differences are 0.11% and 0.41%. These results illustrate
that the role of consistency loss is to make SIC and TPC
collaboratively work better and make their classification results
closer. Furthermore, under the same conditions, the prediction
time of TPC is much larger than that of SIC. Therefore,
we choose the MF joint result of SIC as the final classification
prediction.

7) Comparison of Fusion Strategies: Several classical fea-
ture fusion strategies are selected to replace the MFWF

| Dataset     | Consistency Loss | Time(s) | OA(%)  | AA(%)  | Kappa |
|-------------|------------------|---------|--------|--------|-------|
| SanFrancisco|                  |         |        |        |       |
|             | ×                | 526     | 99.69  | 99.51  | 0.9921|
|             | TPC              | 401     | 97.84  | 97.65  | 0.9117|
|             |                  |         |        |        |       |
|             | SIC              | 528     | 98.16  | 98.69  | 0.9430|
|             | TPC              | 796     | 98.66  | 98.10  | 0.9117|
|             |                  |         |        |        |       |
| Woniupan    |                  |         |        |        |       |
|             | ×                | 381     | 94.59  | 91.80  | 0.9125|
|             | TPC              | 882     | 93.84  | 90.68  | 0.8993|
|             |                  |         |        |        |       |
|             | SIC              | 382     | 95.41  | 93.02  | 0.9290|
|             | TPC              | 669     | 95.01  | 92.77  | 0.9233|
|             |                  |         |        |        |       |
| Flevoland   |                  |         |        |        |       |
|             | ×                | 648     | 95.01  | 91.98  | 0.8425|
|             | TPC              | 1869    | 94.60  | 91.19  | 0.9138|
|             |                  |         |        |        |       |
|             | SIC              | 846     | 95.45  | 92.56  | 0.9477|
|             | TPC              | 1585    | 95.34  | 92.07  | 0.9466|
strategy in MF-STF for comparison, including concatenation fusion, maximization fusion, multiplication fusion, and summation fusion. The corresponding models are represented by Model4, Model5, Model6, and Model7. In addition, based on MFWF, the equal weight summation fusion is also compared, which is denoted by Model8. Fig. 10 summarizes the classification results on three experimental datasets.

Comparing Model4, Model5, Model6, and Model7, it is found in Fig. 10 that the classical fusion strategies for different datasets to achieve the best classification results are different. Specifically, for the three datasets, the optimal classical fusions are multiplication, concatenation, and maximization. Notably, for all datasets, the proposed MFWF strategy can achieve comparable or even better classification results than the optimal classical fusions. For example, for the SanFrancisco dataset, the multiplication fusion holds the highest level of classification performance compared to other fusions. Meanwhile, it is found that the classification results of the proposed MFWF strategy differ the least from the best results.

Furthermore, with the adaptive adjustment of weights, MF-STF is better than Model8. These prove that our MFWF strategy is effective and rather stable, which is conducive to adaptively obtaining more robust classification results.

Furthermore, we also show the learning weights of MFWF in Fig. 11. As mentioned earlier, because the chessboard-like sampling strategy is used, there are two training results, denoted by M1 and M2. As shown in Fig. 11, the contribution of different frequency bands to the final classification decision is different.

D. Classification Results and Comparison

According to the proposed MF-STF, the corresponding models under all single bands and various combinations of different bands are constructed for classification performance comparison. For the two-band SanFrancisco and Woniupan datasets, the band combinations are C&L and S&L, respectively, whereas for the three-band Flevoland dataset, the different band combinations conclude C&L, C&P, L&P, and C&L&P. In addition, some competing works on multisource or PolSAR image classification are employed as comparison algorithms, including Wishart mixture model (WMM) [4], object-based SVM (O-SVM) [24], Stein-sparse representation-based classification (S-SRC) [7], tensor feature-based ANN (TF-ANN) [25], CRPM-Net [2], two-branch CNN (Tb-CNN) [40], depthwise separable convolution-based multitask CNN (DMCNN) [17], and vision transformer (ViT) [23]. To use DMCNN and ViT for MF-PolSAR image classification, the proposed MFWF strategy in this article is adopted to combine bands for the final MF labeling decision.

For a fair comparison, the above DL-based methods all adopt the same sampling strategy as MF-STF. It is noticeable that we only report the results of these comparison methods under the full-band combination.

1) Results on the SanFrancisco Dataset: Table VI shows the quantitative comparison on the SanFrancisco dataset, and Fig. 12 shows the corresponding classification maps for visual evaluation. As shown in Table VI, comparing the results of the two single bands, it can be seen that C-band is superior to L-band in classifying forest and water, while L-band is better than C-band in identifying different kinds of Urbans. By merging C and L, the classification accuracy of all categories is improved. Meanwhile, the OA, AA, and Kappa obtained by C&L are all superior to C and L. These suggest that combining

Fig. 10. Comparison of classification results between the MFWF strategy and other fusion strategies on the three experimental datasets. (a) SanFrancisco. (b) Woniupan. (c) Flevoland.

Fig. 11. Learning weights of different frequency bands by MF-STF on the three datasets. (a) SanFrancisco. (b) Woniupan. (c) Flevoland.

Fig. 12. Classification results of different methods on the SanFrancisco dataset. (a) Ground Truth. (b) MF-STF(C). (c) MF-STF(L). (d) MF-STF(C&L). (e) WMM. (f) O-SVM. (g) S-SRC. (h) TF-ANN. (i) CRP-M-Net. (j) Tb-CNN. (k) DMCNN. (l) ViT.
multiple frequency bands can exploit the complementarity of bands to improve category discrimination. Compared with the existing methods, the proposed MF-STF achieves the best classification performance. It gets over 97% accuracy for all categories. In particular, MF-STF better distinguishes high-density urban, low-density urban, and developed, and the classification accuracy of all three categories exceeds 98%. This demonstrates that the proposed MF-STF can discriminate categories more effectively and maintain a better classification balance among categories.

The classification maps shown in Fig. 12 further verify the effectiveness of the proposed MF-STF on the SanFrancisco dataset. For C-band, there exhibits a serious mix between the classification of high-density urban and low-density urban in Fig. 12(b) (highlighted by white ovals), whereas for L-band, many pixels belonging to forest are misclassified as water in Fig. 12(c) (highlighted by white rectangles). However, these above phenomena are greatly improved by merging dual-band information in Fig. 12(d). As shown in Fig. 12(e)–(l), the existing methods have a weak ability to accurately distinguish between high-density urban and low-density urban. However, our proposed MF-STF can identify them well. In addition, compared with other methods, MF-STF also produces a clearer classification map, which is closer to the ground-truth distribution. In general, the experimental results on the SanFrancisco dataset can demonstrate the classification superiority of the proposed MF-STF.

2) Results on the Woniupan Dataset: For the Woniupan dataset, Table VII and Fig. 13, respectively, show the quantitative and qualitative results. As shown in Table VII, comparing S- and L-bands, the S-band classifies road, building, and forest better than L-band, while for the classification of farmland and bareland, L-band is superior to S-band. When combining dual bands, the classification accuracy of each category obtained by S&L is improved. Furthermore, the classification results of S&L are better than those of S- and L-bands. These results indicate that there is complementarity between the two bands. Compared with traditional methods (WMM, O-SVM, and S-SRC), the DL-based methods (TF-ANN, CRPM-Net, Tb-CNN, DMCNN, ViT, and MF-STF) can greatly improve the classification accuracy of building. The reason may be that the features extracted by DL-based methods have stronger discriminative ability than traditional features and are more conducive to distinguishing similar objects. For all categories except building, MF-STF outperforms other DL-based methods. Notably, the accuracy of Road by MF-STF is over 90%.

Besides, MF-STF obtains the highest values of OA, AA, and Kappa. All of these can prove that the proposed classification method is effective and can achieve more satisfactory classification performance.

Visually, comparing Fig. 13(b) and (c), S-band identifies road well, while L-band can recognize farmland well [high-lighted by blue rectangles in Fig. 13(b) and (c)]. By observing Fig. 13(d), it is found that the classification effect of road and farmland has a certain complementary improvement compared with single bands. In addition, compared with S- and L-bands, the classification result of building in Fig. 13(d) is closer to the ground truth. As shown in Fig. 13(e)–(g), the non-DL methods have worse classification effects on building.
In contrast, TF-ANN, CRPM-Net, Tb-CNN, DMCNN, and ViT have better visual effects in this category because of more discriminative feature extraction. However, for road category with narrow structures, they have insufficient recognition ability [highlighted by white rectangles in Fig. 13(h)–(l)]. It can be seen from Fig. 13(d) that the proposed MF-STF can identify road well and preserve its fine structure better. In addition, the visual result of MF-STF is closest to the ground truth.

3) Results on the Flevoland Dataset: For the Flevoland dataset, Table VIII quantitatively shows the performance comparison. Besides, Fig. 14 shows the relevant classification results. As shown in Table VIII, L-band achieves higher classification accuracy than C- and P-bands for most categories. C-band is more suitable to identify grass, potato, and beans. P-band is superior to other bands in classifying wheat. The main reason can be explained by the band difference.
As previously analyzed, compared with C- and P-bands, the L-band microwave has a certain penetration ability. It can sense the different responses of the branches under canopies to better distinguish similar crops. Meanwhile, these experimental results reflect that different bands are suitable for identifying different ground objects. The dual-band combinations (C&L, C&P, and L&P) all yield better accuracy than the single band. Notably, the C&L&P combination gets the best accuracy than single-band and other combinations. Except for beans, fruit, and lucerne, the accuracy of other categories in the C&L&P combination is better than that of single band. These illustrate that merging multiple bands indeed improves the classification performance under single-band conditions. For the C&L&P combination, MF-STF exhibits higher OA, AA, and Kappa than existing methods. In addition, compared with other DL-based methods (TF-ANN, CRPM-Net, Tb-CNN, DMCNN, and ViT), MF-STF achieves the best classification accuracy in most categories. These results show that our designed network improves feature discrimination, thereby better distinguishing similar categories.

Visually, as shown in Fig. 14(b)–(d), many pixels in single-band classification maps are misclassified. For example, wheat and barley are heavily mixed with other categories [highlighted by black rectangles in Fig. 14(b)–(d)]. This phenomenon is alleviated in Fig. 14(e)–(h) because of the efficient use of multiband features. In particular, the classification map of C&L&P has smoother homogeneous areas and better connectivity. As shown in Fig. 14(i)–(k), WMM, O-SVM, and S-SRC have good regional category consistency, but there are obvious regional misclassifications (highlighted by black ovals). For TF-ANN, it is clear that many areas are misclassified as flax in Fig. 14(l) (highlighted by white ovals). In addition, in Fig. 14(m), CRPM-Net misclassifies many pixels as lucerne (highlighted by white ovals). As shown in Fig. 14(p), the misclassification of wheat and barley is more serious (highlighted by white ovals). Compared with TF-ANN, CRPM-Net, and ViT, Tb-CNN and DMCNN have better visual effects and the misclassification points are greatly reduced. Despite this, there are still some unreasonable distributions in Tb-CNN and DMCNN, which are away from the real ones. For example, building pixels appear in other areas and are mixed with fruit pixels in Fig. 14(n) (highlighted by black rectangles), and some pixels belonging to grass are misclassified in Fig. 14(o) (highlighted by black rectangles). Compared with existing methods, MF-STF obtains the result closer to the ground truth and takes the lead in terms of boundary position recognition and region label consistency. In summary, according to Table VIII and Fig. 14, it can be concluded that the proposed MF-STF outperforms other methods on the Flevoland dataset.

To sum up, the experimental results on three MF datasets show that merging MF information can eliminate the classification inaccuracy under single band and is beneficial to distinguishing similar categories. In addition, compared with existing methods, the proposed MF-STF achieves remarkable performance improvement, which demonstrates that our designed modules and network are effective. With the help of CIFE, as well as jointly considering local and nonlocal spatiality, MF-STF can capture and utilize more discriminative information, thereby obtaining more satisfactory performance.

### IV. Conclusion

This article proposes MF-STF for MF-PolSAR image classification, which aims to fully mine and leverage the complementarity of bands and combine local and nonlocal spatial information to improve classification accuracy. In MF-STF, based on the band-specific semantic representation, the proposed CIFE module explicitly models the deep semantic correlation among bands. It realizes the interactive fusion and enhancement of interband information, thereby making full use of the complementarity of bands to improve the accuracy of ground object classification. In addition, MF-STF adopts the GraphSAGE model to dynamically capture the representation of nonlocal topological relations between samples. In this way, the local and the nonlocal spatial information are captured simultaneously, which further improves the robustness of classification results. Moreover, based on the total loss,
the proposed MF-WF strategy adaptively updates weights to flexibly fuse inferences from different bands for the final MF joint classification decision. Experiments on three measured MF-PolSAR datasets show that the proposed modules are effective in improving classification performance. In addition, the proposed MF-STF can more effectively capture and utilize the complementarity of bands to eliminate the classification inaccuracy under SF conditions. Furthermore, MF-STF can combine local and nonlocal spatial information to obtain more accurate results than other related state-of-the-art classification methods.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions that greatly strengthened this article.

REFERENCES

[1] J. S. Lee and E. Pottier, Polarimetric Radar Imaging From Basic to Application. New York, NY, USA: CRC Press, 2011.
[2] D. G. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
[3] C. Szegedy et al., “Going deep with widths,” in Proc. Neural Inf. Process. Syst., 2015, pp. 1–9.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “One weird trick for improving generalization,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 3333–3341.
[5] I. Goodfellow et al., “Deep learning,” in Deep Learning. MIT Press, 2016, pp. 1–102.
[6] G. E. Hinton and O. Vinyals, “Distributed representations of words and phrases and their compositionality,” in Adv. Neural Inf. Process. Syst., 2013, pp. 3111–3119.
[7] J. Levy et al., “Leveraging prosody to improve statistical models of speech,” in Proc. Interspeech, 2014, pp. 4129–4133.
[8] D. Y. Povey et al., “The Kaldi speech recognition toolkit,” in Proc. Workshop Open Source Speech, 2011.
[9] S. Ren et al., “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Neural Inf. Process. Syst., 2015, pp. 91–99.
[10] R. Girshick et al., “Fast R-CNN,” in Proc. Neural Inf. Process. Syst., 2015, pp. 1440–1448.
[11] S. Ren et al., “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Neural Inf. Process. Syst., 2015, pp. 91–99.
[12] R. Girshick et al., “Fast R-CNN,” in Proc. Neural Inf. Process. Syst., 2015, pp. 1440–1448.
[13] S. Ren et al., “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Neural Inf. Process. Syst., 2015, pp. 91–99.
[14] R. Girshick et al., “Fast R-CNN,” in Proc. Neural Inf. Process. Syst., 2015, pp. 1440–1448.
[38] Y. Gao et al., “Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5512615, doi: 10.1109/TGRS.2021.3097093.

[39] W. Liang, Y. Wu, M. Li, and Y. Cao, “High-resolution SAR image classification using context-aware encoder network and hybrid conditional random field model,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 8, pp. 5317–5335, Aug. 2020.

[40] 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.

Yice Cao (Member, IEEE) received the Ph.D. degree from the Remote Sensing Image Processing and Fusion Group, School of Electronic Engineering, Xidian University, Xi’an, China, in 2022. She is currently a Lecturer at the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei, China. Her primary research focuses on intelligent interpretation of remote sensing images and deep learning.

Yan Wu (Member, IEEE) received the B.S. degree in information processing and the M.A. 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 Post-Doctoral Fellow with the National Key Laboratory of Radar Signal Processing, Xidian University. Since 2005, she has been a Professor with the Department of Electronic Engineering, Xidian University. She has authored or coauthored more than 80 technical articles. 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.

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.

Mingjie Zheng received the Ph.D. degree in engineering from the Institute of Electronics, Chinese Academy of Sciences (IECAS), Beijing, China, in 2003. In 2003, she joined the Department of Space Microwave Remote Sensing System, IECAS, engaged in spaceborne synthetic aperture radar (SAR) system design and simulation analysis, during which she presided some research projects about spaceborne SAR technology. She is currently a Research Fellow and a Doctoral Supervisor with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include spaceborne SAR ground moving target indicator (GMTI) technology, spaceborne SAR full-link system design, and spaceborne SAR dual-frequency interferometric technology and polarimetric technology.

Peng Zhang received the B.S. degree in electronic and information engineering and 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 at the National Key Laboratory of Radar Signal Processing, Xidian University. His main research interests are synthetic aperture radar (SAR) image interpretation and statistical learning theory.

Jili Wang received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2013, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, in 2019. In 2019, he joined the Spaceborne Microwave Remote Sensing System Department, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include the development of synthetic aperture radar (SAR) image application techniques, SAR interferometry techniques (InSAR), time-series InSAR processing techniques, and their applications.