Dual-Branch Feature Interaction Network for Coastal Wetland Classification Using Sentinel-1 and 2

Mingming Xu, *Member, IEEE*, Mingwei Liu, Yanfen Liu, Shanwei Liu, and Hui Sheng

***Abstract*—The combination of multispectral image (MSI) and synthetic aperture radar (SAR) data has made certain progress in coastal wetland classification. How to realize the interactive fusion between the two data and make full use of their fusion characteristics becomes challenging. However, the existing joint classification methods neglect interaction information between features and underutilize fusion features. Therefore, this article proposes a dual-branch feature interaction network (DFI-Net) that joins MSI and SAR data for coastal wetland classification. The dual-branch independent structure of 3DCNN processing MSI and 2DCNN processing SAR is designed, which can effectively capture spectral–spatial features and polarization features. In addition, we develop two novel modules. The feature interaction fusion block is designed to enhance the complementarity between the features of the two kinds of data. This block employs a cross-agent attention mechanism to realize effective interaction between MSI and SAR features and adaptive fusion of contextual information from the two branches. Finally, a plug-and-play module channel–spatial transformer encode (CSTE) is proposed to improve the utilization rate of interactive fusion data. The CSTE utilizes two parallel transformers to deeply mine information in interactive fusion data and explore channel–spatial features across all dimensions to the maximum extent possible. The classification experiment is conducted on the Yellow River Delta coastal wetland dataset. The experimental results show that the overall accuracy of DFI-Net reaches 97.03%, which outperforms the performance of other competitive approaches. The effectiveness of DFI-Net provides a reference method for combining MSI and SAR for coastal wetland classification.**

***Index Terms*—Classification, coastal wetland, data interactive fusion, multispectral image (MSI), synthetic aperture radar (SAR).**

I. Introduction

Wetlands, alongside oceans and forests, constitute the three major ecosystems globally, renowned as the kidneys of the Earth, natural reservoirs, and treasuries of species. Coastal wetlands, situated in the transition zone between sea and land [1], play a crucial role in providing ecosystem services, offering habitats for wildlife, safeguarding coastlines, and promoting global economic development [2], [3], [4]. However, due to natural and anthropogenic factors, coastal wetlands are facing challenges, such as a reduction in scale, degradation of ecological carrying capacity, and decline in habitat quality [5]. Therefore, the protection of coastal wetlands is urgent, and classification is an essential prerequisite for the better protection of coastal wetland ecosystems.

The traditional wetland survey methods primarily rely on manual interpretation and field surveys, which are characterized by low efficiency, high costs, and long time cycles [6]. It is not easy to meet the monitoring and management needs of large-scale wetlands using these methods. Remote sensing (RS) techniques have overcome the limitations of traditional methods, providing significant convenience for wetland classification [7]. Especially, multispectral image (MSI) has been widely applied to wetland classification due to its characteristics of high spatial resolution and wide spatial coverage [8], [9], [10]. However, using MSI alone for coastal wetland classification will face some difficulties. On the one hand, the highly fragmented landscape pattern of coastal wetlands leads to significant changes in the shape and scale of land features, which increases the interclass variability and decreases the intraclass similarity. On the other hand, some vegetation classes have overlapping spectral reflectance during peak growing seasons. Similar spectral features make it difficult to accurately identify different vegetation types [11], [12].

Fortunately, synthetic aperture radar (SAR) is an active imaging sensor that contains abundant information, such as phase, intensity, and polarization [13], [14]. Because of the different Earth surface information from MSI, SAR is also often used for coastal wetland classification and mapping [15], [16]. Furthermore, the ability of SAR signals to penetrate through vegetation canopies gives them an advantage over optical sensors [17], [18]. Therefore, considering that SAR can be an effective supplementary data to MSI, the combination of MSI and SAR data has become a widely adopted method for coastal wetland classification. In addition, the high spatiotemporal resolution Sentinel-1 SAR (S1) and Sentinel-2 MSI (S2) collected by the European Space Agency provide free data support for wetland classification and monitoring [19]. Researchers have done a lot of research on combining S1/2 (S1 and S2) data for wetland classification [20], [21], [22].

Manuscript received 15 May 2024; revised 18 July 2024; accepted 3 August 2024. Date of publication 8 August 2024; date of current version 26 August 2024. This work was supported in part by the Open Fund for Observation and Research Station of Bohai Strait Eco-Corridor under Grant BH202301, in part by the Fundamental Research Funds for the Central Universities under Grant 24CX07005A, and in part by Shandong Provincial Colleges and Universities Youth Innovation Technology Support Program under Grant 2023KJ068. *(Corresponding author: Yanfen Liu.)*

Mingming Xu, Mingwei Liu, Shanwei Liu, and Hui Sheng are with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China (e-mail: xumingming@upc.edu.cn; lmw991217@163.com; shanweiliu@163.com; sheng@upc.edu.cn).

Yanfen Liu is with the Observation and Research Station of Bohai Strait Eco-Corridor, MNR, Qingdao 266061, China, and also with the Dongying Marine Development Research Institute, Dongying 257091, China (e-mail: liuyanfen0319@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3440640

Typical machine learning methods, such as decision tree, support vector machine, random forest, and Bayesian optimized tree, were commonly used in the early stages of research for wetland classification combining MSI and SAR data [23], [24], [25]. Feature extraction and feature selection are also major challenges in multisource RS wetland classification. In response to this challenge, many scholars have carried out feature optimization research to improve the accuracy of wetland classification based on traditional machine learning methods [26], [27]. Some machine learning algorithms have been applied to coastal wetland classification and worked well [28]. However, traditional machine learning methods often rely on manually designed features. Different methods will choose different features, which can lead to bias in feature selection. As a result, the model overly relies on these specific features, performs well on training data, has poor generalization ability on new data, and increases the risk of overfitting. In addition, machine learning is prone to ignoring the correlation between features, which has obvious limitations when dealing with features with complex relationships.

In recent years, the application of deep learning to combine multisource data for wetland classification has gradually become a research hotspot. Deep learning has strong feature learning capabilities. When dealing with complex wetland environments, subtle differences and complex patterns that are difficult to recognize by humans can be captured without manually designing features [29]. In addition, deep-learning models exhibit good generalization ability. It is guaranteed to maintain high accuracy on previously unseen test data after training. Deep-learning methods have been proven to perform well in swamp vegetation mapping tasks [30]. At present, the joint MSI and SAR wetland classification model based on deep learning can be divided into single-branch network [31], [32], [33] and multibranch network [34], [35].

Convolutional neural network (CNN) is one of the typical single-branch classification algorithms. DeLuncey et al. [36] successful discriminability of the wetland categories Unet-CNN using MSI and SAR images. However, the deep convolutional structure consumes a lot of time on CNN and CNN cannot capture rich global information limited by the convolution kernel. Therefore, the researchers explored multimodel combined single-branch deep-learning architectures, including the model integrating AlexNet and generative adversarial network (GAN) and the combination of CNN and vision transformer [37], [38]. Compared with CNN alone, multimodel combined network can make full use of the advantages of different networks and improve the accuracy of wetland classification, which is particularly important for capturing complex spatial relationships and heterogeneous features in wetland classification. What is more, the multisource transformer under a single-branch network framework is an excellent tool for drawing wetland maps [39], [40]. The fractional Fourier image transformer also provides the latest idea for multisource data joint classification of wetlands [41]. Unfortunately, the single-branch classification algorithms do not consider the difference between MSI and SAR data, and it is not reasonable to process data with different information in the same way. Therefore, multibranch networks that adopt different processing methods for different data are gradually being proposed. Research has shown that the multibranch structure has good application capabilities and performs well in wetland mapping. The dual-branch network is an effective framework for increasing the possibility of wetland classification using both MSI and SAR data [42]. If there are other types of data to assist in classification besides MSI and SAR data, a three-branch network can be used. Recent articles have demonstrated that using three-branch networks for wetland classification has great potential [43], [44]. Whether it is a dual-branch or three-branch network design, their commonality is that they allow for simultaneous processing of MSI and SAR data. They can also utilize the unique functionality of each branch to extract features from different data to capture heterogeneous information. Specifically, the focus of the above methods is to design multibranch MSI and SAR feature extraction models. In the feature fusion stage, only the extracted effective features are directly stacked or simply concatenated, which does not consider the interaction information between MSI and SAR features. However, feature interaction can obtain more discriminant fusion features, which is also important for wetland classification using MSI and SAR data. One aspect of feature interaction is how to design a practical feature interaction fusion module to ensure features after interaction fusion possess higher information richness. Another aspect is how to realize the deep mining and utilization of interactive fusion features.

To solve the above problems, we propose a dual-branch feature interaction network (DFI-Net) for joint coastal wetland classification using MSI and SAR data. First, double-branch networks are used to simultaneously extract MSI space-spectral features and SAR polarization features. Second, cross-agent attention is introduced to design a feature interaction fusion block (FIFB), which is used to cross and fuse contextual information from both branches and realize the complementary advantages between MSI and SAR data. Then, we develop an efficient channel–spatial transformer encoder (CSTE) module. Through two parallel transformer encode branches, the module deeply mines the channel–spatial information in the interactive fusion data in a combinatorial manner. Finally, the deep features extracted from the two branches are combined into the classifier to obtain the classification result. The main contributions of DFI-Net are summarized as follows.

1) In view of the difference between MSI and SAR data, this article designs a DFI-Net that can process different types of data simultaneously. DFI-Net readjusts the feature interaction mechanism and deep information mining structure to better realize the feature complementarity and information exchange between MSI and SAR data.
2) An FIFB is proposed to perform the information interaction of MSI and SAR features. This module takes the features extracted from two CNN branches as inputs and achieves cross-agent attention between MSI and SAR data by generating self-attention on the feature map. FIFB promotes the complementary advantages of MSI and SAR data, which improves the expression ability of fused features.

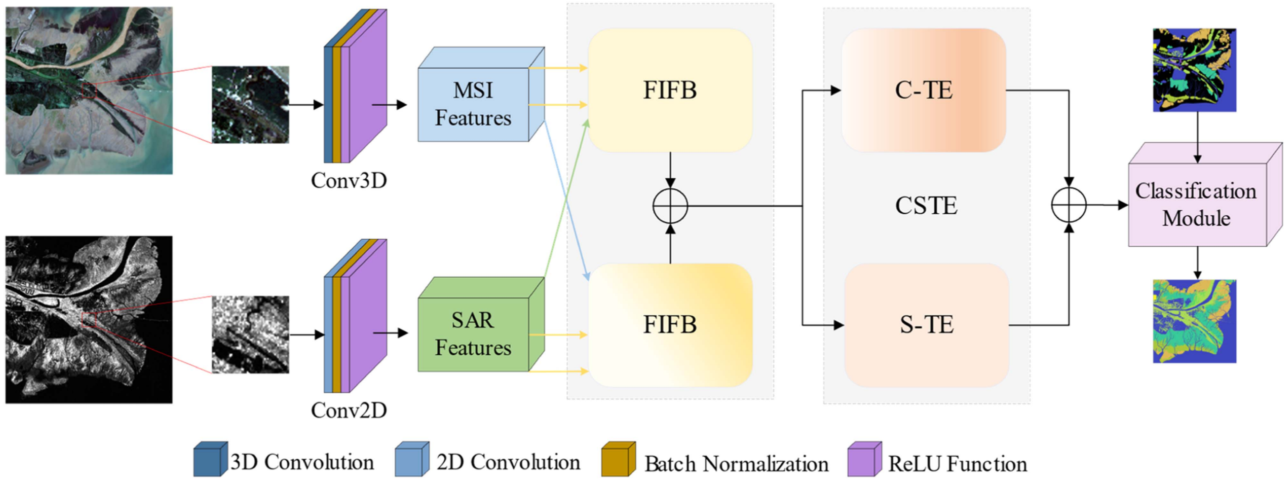


Fig. 1. Illustrates the proposed DFI-Net, including the preliminary feature extraction module, FIFB, and deep mining module.

3) In order to improve the utilization of interactive fusion data, a CSTE module is developed to mine its channel–spatial characteristics deeply. The attention weight is generated by the channel and spatial position information, and then the convolutional layer is fed, respectively, to generate deep channel–spatial features. The introduction of rich deep features improves discriminability and classification accuracy. In addition, the designed CSTE is flexible and effective. It can be used as a plug-and-play module.

The rest of this article is organized as follows. Section II presents the details of the proposed method DFI-Net. Section III introduces the study area and datasets. In Section IV, the experimental results are analyzed. Section V provides the discussion. Finally, Section VI concludes this article.

II. Proposed Method

The overall network of DFI-Net for coastal wetland classification using MSI and SAR data is shown in Fig. 1. It consists of four parts.

1) The 3D–2DCNN, preliminary feature extraction for MSI and SAR.
2) FIFB achieves complementary advantages between MSI and SAR data through feature interaction and fusion.
3) CSTE for deep mining of channel–spatial information in interactive fusion data.
4) Classifier, which inputs synthetic features into the classifier to obtain a classification map.

A. MSI and SAR Feature Extraction Via CNNs

CNN is one of the most famous deep-learning algorithms, which can automatically learn local features and texture information in RS images. It is very effective for 2-D RS data. In addition, 3DCNN has been shown to have a unique advantage in processing 3-D RS data by utilizing spectral–spatial information simultaneously. Therefore, in the proposed network, 3DCNN and 2DCNN double-branch structures are used to extract spectral-space features from MSI data and polarization features from SAR data, respectively.

As shown in Fig. 1, the proposed dual-branch CNN network has a simple and shallow structure. Extract discriminative spatial and spectral features from MSI data using 3DCNN. Each MSI patch cube $\mathbf{X}_M^P$ of size $s \times s \times b$ is used as the input training data for the Conv3-D layer. In the Conv3-D layer, the size of the convolution kernel is set to 64@$1 \times 1 \times 1$. After the convolution operation, 64 2-D feature maps can be generated.

Unlike MSI processing, 2DCNN is used to extract polarization features from SAR data. Each SAR patch cube $\mathbf{X}_S^P$ of size $s \times s$ serves as the input to the Conv2-D layer. In this layer, the size of the convolution kernel is set to 64@5×5. To regularize and accelerate the training process, the batch normalization layer and rectified linear unit layer are continuously applied after the convolutional layer.

B. Feature Interactive Fusion Block

The complementary information between MSI and SAR can be captured through feature interaction. Cross attention can interactively process different types of data [45]. Agent attention has the advantages of both linear complexity and high expressiveness [46]. Inspired by the above two types of attention, we design a new cross-agent attention called FIFB, as shown in Fig. 2. Specifically, it consists of two steps. The first step is to generate agent token A using the features of one type of data as the main feature and perform attention calculation between key K and value V generated by the features of another type of data. This step aggregates two heterogeneous features to obtain the agent feature V_A. The second step is to use A as the key and V_A as the value to perform the second attention calculation using the query Q generated by the main features. In our network, there are two FIFBs. The features of MSI and SAR data are sequentially used as the main features of the interaction fusion module. After obtaining the two interaction fusion results, they are added together to get the final interaction fusion feature. This module enables efficient interaction between MSI and SAR

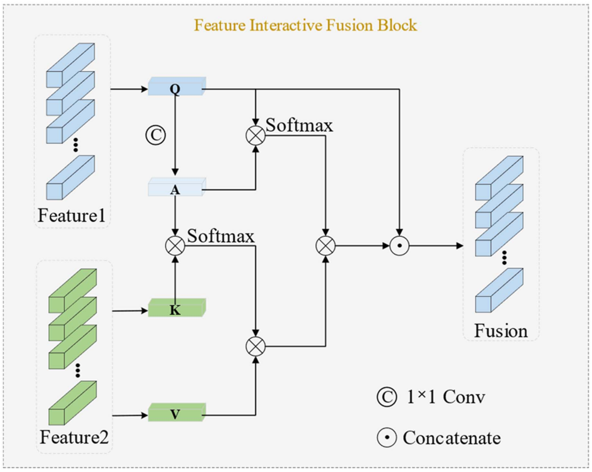


Fig. 2. Overview of the FIFB.

features and adaptive fusion of the two branches. It is worth noting that agent attention has been applied for the first time in wetland classification tasks. Next, we will provide a detailed introduction to the implementation process of FIFB.

First, we select F_1 as the main feature and obtain Q through a linear project. Then, the pooling strategy is further used for Q to obtain A. The K and V are obtained through the linear project using another feature F_2. The process can be represented as follows:

$$Q = W_q\ F_1,\ K = W_k\ F_2,\ V = W_v\ F_2 \tag{1}$$

$$A = \ \text{pool}\,(Q) \tag{2}$$

where W_q, W_k, and W_v are the learnable weights, and pool$(\cdot)$ represents the pooling operations.

Next, we feed Q, A, K, and V into the agent attention mechanism to calculate the interactive fusion result T. Agent attention consists of two parts of operations: agent aggregation and agent broadcast. The agent aggregation operation treats the A as a query and performs attention calculations among A, K, and V to get the V_A. In agent broadcast operation, A is used as key and V_A as value, and Q is used for the second attention calculation to obtain the output A_T. The essence of this process is that the newly defined agent token A acts as Q, aggregates global information from K and V, and then broadcasts it back to Q. Finally, the A_T is connected with the input main feature F_1 to obtain the T. The calculation process is expressed as follows:

$$T = \text{Cat}\left(\sigma\left(QA^T\right)\sigma\left(AK^T\right)\ V, F_1\right) \tag{3}$$

where $\sigma(\cdot)$ represents the softmax function, and Cat$(\cdot)$ represents the connection of each channel.

We obtain T_S^M and T_M^S for MSI and SAR data through FIFB, respectively. Add them together to get the final interaction fusion feature T_{end}, as shown in the following formula:

$$\left.\begin{matrix} F_{\text{MSI}}\ as\ F_1 \\ F_{\text{SAR}}\ as\ F_2 \end{matrix}\right\} \Rightarrow T_S^M$$

$$\left.\begin{matrix} F_{\text{SAR}}\ as\ F_1 \\ F_{\text{MSI}}\ as\ F_2 \end{matrix}\right\} \Rightarrow T_M^S$$

$$T_{\text{end}} = T_S^M + T_M^S. \tag{4}$$

C. CSTE Module

To further improve the utilization rate of T_{end}, we design a new form of transformer encoder module-CSTE, inspired by channel and position attention [47]. The detailed structure is shown in Fig. 3. CSTE consists of two parallel channel transformer encoder and spatial transformer encoder, which are used to mine deep channel–spatial information of interactive fusion data. The first innovation is to introduce a deep convolutional layer between attention through residual connection. The second innovation is to integrate two transformer encoders into a unified module. CSTE first calculates the weight distribution of interactive fusion data through attention, in order to capture global contextual information. This step ensures information integrity and contextual relevance during the deep feature mining process. Then, add a convolutional layer after the attention mechanism. The embedding of the convolutional layer enhances the ability to represent local details. The parallel connection structure of CTE and STE can simultaneously focus on spatial and channel information, and the extracted deep features are comprehensive and accurate. Specifically, the design of the CSTE approach has strong flexibility and can be seamlessly integrated into the existing deep-learning models.

1) Spatial Transformer Encode: As shown in Fig. 3, the spatial transformer encoding (STE) consists of a spatial position attention module (PAM) and deep convolution. A residual skip connection is designed before the PAM block and deep convolution. Precisely, PAM attention mechanisms combined with convolution operations can capture global–local information between different positions to enhance feature expression capabilities. Residual connections help the model train more stably. The following is the implementation of STE.

The interaction fusion feature $T_{\text{end}} \in \mathbb{R}^{C\times H\times W}$ generates two new feature maps E and F through a convolutional layer, where $\{E,\ F\} \in \mathbb{R}^{C\times H\times W}$. Then, through the reshape operation, the E and F shapes are changed to $\mathbb{R}^{C\times N}$, where $N = H \times W$, representing the number of pixels. After matrix multiplication of E and F, the softmax layer is applied to obtain the spatial attention map $S \in \mathbb{R}^{N\times N}$. The calculation process is represented as follows:

$$S_{ji} = \frac{\exp\left(E_i \cdot F_i\right)}{\sum_{i=1}^{N} \exp\left(E_i \cdot F_j\right)} \tag{5}$$

where S_{ji} represents the influence of the ith position in a pixel on the jth position. If the features of two pixels are similar, this value represents their correlation strength.

While generating the E and F feature maps, input the T_{end} into the convolutional layer to generate a new feature map $D \in \mathbb{R}^{C\times H\times W}$. And reshape it as a tensor of $\mathbb{R}^{C\times N}$. After obtaining the new tensor, perform matrix multiplication with the transpose of S and reshape the result as the feature map of $\mathbb{R}^{C\times H\times W}$.

After calculating the feature map, it is multiplied by learnable parameters α and added to the residual of the T_{end} to obtain the

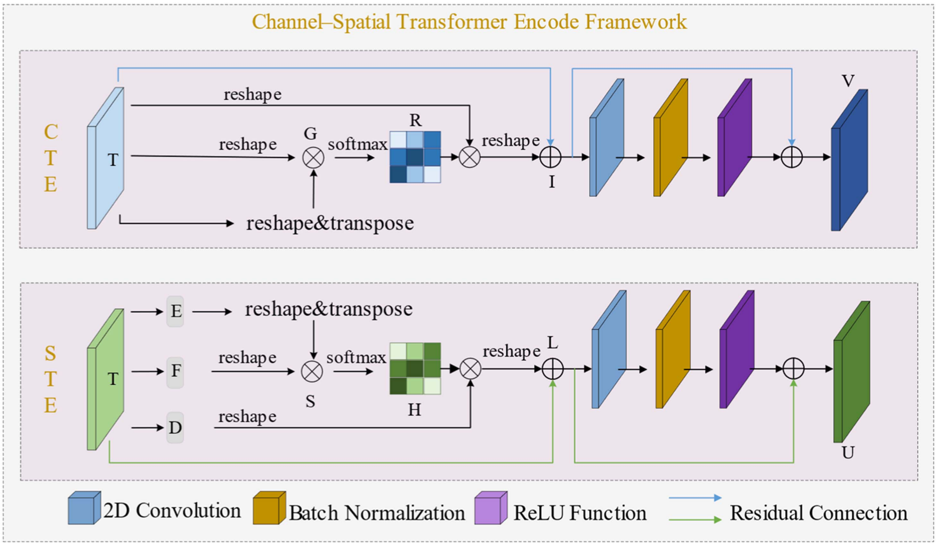


Fig. 3. Detailed structure of the CSTE module.

output $L \in \mathbb{R}^{C\times H\times W}$, which is represented as follows:

$$L_j = \alpha \sum_{i=1}^{N} (S_{ji} D_i) + T_{\text{end}j}. \tag{6}$$

Finally, a deep convolution layer is introduced and connected with L residual to obtain the final output $U \in \mathbb{R}^{C\times H\times W}$, which is expressed by the equation:

$$U = \text{Conv2D}\,(L) + L \tag{7}$$

where $\text{Conv2D}(\cdot)$ represents the 2-D convolution operation.

2) Channel Transformer Encode (CTE): The CTE consists of a channel attention module (CAM) and deep convolution. As shown in Fig. 3, residual skip connections are also designed before the CAM block and deep convolution. Unlike PAM, CAM directly calculates the channel attention map $G \in \mathbb{R}^{C\times C}$ in $T_{\text{end}} \in \mathbb{R}^{C\times H\times W}$.

First, the T_{end} is reshaped from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{C\times N}$. Then, vector operations are performed on T_{end} and its transpose. At last, the softmax activation function is used to obtain the channel attention map $G \in \mathbb{R}^{C\times C}$. The process is represented as follows:

$$G_{ji} = \frac{\exp\left(T_{\text{end}i} \cdot T_{\text{end}i}\right)}{\sum_{i=1}^{C} \exp\left(T_{\text{end}i} \cdot T_{\text{end}j}\right)} \tag{8}$$

where G_{ji} represents the influence of layer i feature map channel ith on layer j feature map channel jth.

Next, vector operations are performed between G and the transpose of T_{end}, followed by reshaping the result to $\mathbb{R}^{C\times H\times W}$ using reshape. The resulting tensor is combined with a scaling parameter β and then undergoes elementwise residual summation with the T_{end} to obtain the output $I \in \mathbb{R}^{C\times H\times W}$. The calculation process is represented as follows:

$$I_j = \beta \sum_{i=1}^{c} \left(G_{ji} \cdot T_{\text{end}j}\right) + T_{\text{end}j}. \tag{9}$$

Finally, we introduce a deep convolutional layer and make residual connections with I to obtain the final output $V \in \mathbb{R}^{C\times H\times W}$. It can be expressed as follows:

$$V = \text{Conv2D}\,(I) + I. \tag{10}$$

D. Classification Module

The output of the CSTE is fed into the multilayer perceptron (MLP) layer. The MLP has two linear layers with Gaussian error linear unit (GELU) operations implemented by the fully connected (FC) layer. GELU in MLP is a standard function defined as follows:

$$\text{GELU}\ (V) = V\Phi\ (V) = \frac{V}{2}\left[1 + \text{erf}\left(\frac{V}{\sqrt{2}}\right)\right] \tag{11}$$

where $\Phi(V)$ represents the standard Gaussian cumulative distribution function, $\text{erf}(V) = \int_0^V\ e^{-t^2} dt$. The MLP module is summarized as follows:

$$\text{MLP}\ (V) = \text{FC}_2\ (\text{GELU}\,(\text{FC}_1\,(V)))\,. \tag{12}$$

After MLP, a global average pooling and FC layer are used to obtain the final classification result.

III. Study Area and Datasets

A. Study Area

The Yellow River Delta is located in the northeast of Shandong Province, China, between Bohai Bay and Laizhou Bay. It is an alluvial plain formed by the sediment of the Yellow River in the Bohai depression. The Yellow River Delta boasts the world's largest, best-preserved, and youngest coastal wetland ecosystem. This study selects the estuary of the Yellow River Delta as the research area, as shown in Fig. 4.

This area is located at midlatitude and belongs to a temperate continental monsoon climate with four distinct seasons. The average annual temperature is about 12.9 °C, and the average

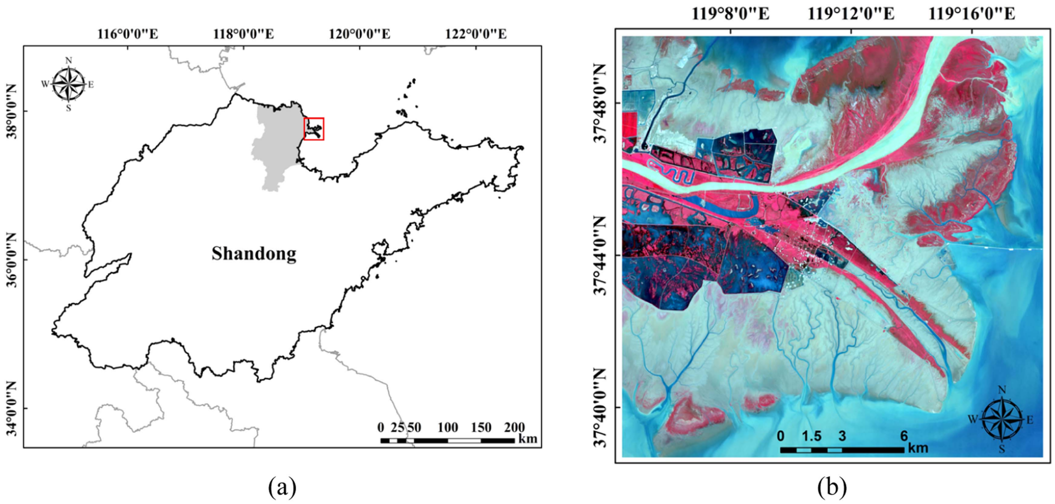


Fig. 4. (a) Overview of the study area. (b) Pseudocolor image for MSI (R, Band8; G, Band4; and B, Band3) in Yellow River estuary.

yearly precipitation is 560 mm. In the study area, vegetation coverage is high, and wetland vegetation types are abundant. The main wetland vegetation types are natural vegetation, such as tamarix chinensis, suaeda salsa, phragmites communis, willow forests, and exotic species spartina alterniflora. Different types of vegetation may have similar spectral characteristics or spatial distribution, which makes the identification and classification of vegetation communities relatively difficult.

B. Datasets Description

1) Sentinel-1/2 Datasets and Preprocessing: The sentinel series data are powerful and freely available RS data from the existing satellite images, known for their high spatial resolution, good spectral quality, and easy accessibility. S1 and S2 are the source guarantees for SAR and MSI data in wetland classification tasks and can be downloaded for free from the European Space Agency's Copernicus Open Access Center.

S1 operates at a *C*-band wavelength (centered around 5.55 cm) with an orbital altitude of 700 km and a revisit period of 6 days. We acquired the S1 interference wide field ground-area detection level-1 backscattering coefficient product for 2021. First, S1 data are preprocessed using the SNAP platform, which includes track correction, thermal noise removal, radiometric calibration, refined Lee filtering, terrain correction, and cropping. Second, the difference and ratio of VV and VH bands are calculated.

The S2 imaging has a width of 290 km and covers 13 spectral bands, which are divided into visible light, near infrared, and shortwave infrared. We got S2 MSI level-2A products for 2021. First of all, we perform resampling and cropping preprocessing operations on the S2 data. Second, ten widely used bands for wetland classification were selected according to research needs, namely B2, B3, B4, B5, B6, B7, B8, B8a, B11, and B12 [48], [49]. After resampling the MSI and SAR images to a spatial resolution of 10 m × 10 m, we performed georeferencing to unify the projection coordinate system into WGS-84 UTM Zone 50N. The S1 and S2 data information used in the experiment is shown in Table I.

2) Ground Truth (GT): In August 2023, the research group went to the Yellow River Delta to carry out a field investigation. The location of accessible sample points is recorded using the global positioning system (GPS) and the photographs of the site are taken. In areas inaccessible to personnel, UAR aerial photography is used to conduct surveys and identify vegetation species through visual interpretation. Eight types of wetland cover were divided into salt soil, natural willow forest, spartina alterniflora, reed, tamarisk, cultivated land, water body, and tidal flat. At the same time, by combining historical data and Google Earth image, manual interpretation of MSI image is performed to complete the drawing of GT. The S1, S2, and GT of the study area together constitute the Yellow River delta coastal wetland dataset (YRCWD), as shown in Fig. 5.

The research area is 2048 × 2048 pixels. After removing background pixels, 2 440 523 samples are retained. Randomly select 0.5% of the samples from each category for training and validation, with a 1:1 ratio of training and validation. The remaining samples are used for testing. Detailed information about the samples used for training, validation, and testing for each class can be found in Table II.

IV. Experimental Results and Analysis

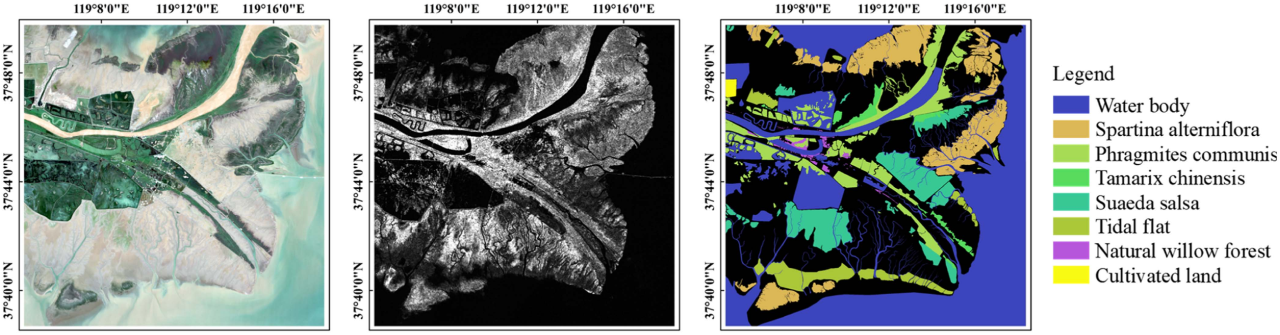
A. Experimental Setup

1) Implementation Details: To ensure experimental fairness, all methods are implemented on the deep-learning framework of PyTorch 1.13.1, and use NVIDIA GeForce RTX 3060 GPU, Intel Core i5-12490F CPU, and 16-GB RAM platform for training. For the training phase, the batch size and the number of training periods are set to 128 and 200, respectively. The loss function uses CrossEntropy. The Adam algorithm is chosen as the initial optimizer to optimize the network.

2) Evaluation Metrics: To evaluate the classification performance of the proposed network and other existing models, calculate three commonly used evaluation metrics, including overall

TABLE I
DETAILED INFORMATION ON S1 AND S2 DATA

Satellite Data	Sensors	Bands	Time
Sentinel-1B	C-SAR	VV, VH, VV-VH, VV/VH	2021.7.23
Sentinel-2B	MSI	B2–B8a, B11–B12	2021.7.25



(a) (b) (c)

Fig. 5. YRCWD. (a) RGB illustration of S2 image (MSI). (b) Grayscale VH band illustration of S1 image (SAR). (c) GT.

TABLE II
NUMBER OF TRAINING, VALIDATION, AND TESTING SAMPLES FOR THE YRCWD

No.	Classes	Train.	Val.	Test.
1	Waterbody	3470	3471	1381421
2	Spartina alterniflora	865	864	344071
3	Phragmites communis	662	661	263322
4	Tamarix chinensis	107	107	42494
5	Suaeda salsa	626	626	249079
6	Tidal flat	306	307	122022
7	Natural willow forest	43	43	17099
8	Cultivated land	22	22	8813
	Total	6101	6101	2428321

TABLE III
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON THE YRCWD

Class	Stack-3DCNN	Stack-RSSAN	Stack-ABLSTM	Stack-Speformer	Stack-GAHT	HCTNet	ExViT	Proposed
1	98.08	98.31	97.69	98.04	98.53	**98.62**	98.56	98.41
2	94.29	96.44	94.78	94.74	96.96	96.81	96.49	**97.79**
3	91.26	93.49	92.02	89.13	93.94	94.53	93.79	**94.69**
4	71.71	86.48	80.64	84.13	90.19	89.17	88.61	**92.02**
5	92.45	94.46	91.99	93.23	95.12	94.05	93.92	**96.14**
6	89.07	92.13	91.13	90.57	91.21	**95.37**	94.15	94.52
7	0.65	21.64	9.61	29.12	42.68	12.18	24.89	**51.13**
8	83.67	90.21	85.11	87.21	91.68	93.46	92.89	**94.00**
OA(%)	94.58	96.04	94.78	94.97	96.53	96.49	96.34	**97.03**
AA(%)	77.65	84.14	80.37	83.27	87.54	84.27	85.41	**89.84**
K×100	91.40	93.72	91.76	92.04	94.50	94.44	94.19	**95.30**

accuracy (OA), average accuracy (AA), and kappa coefficient (K). For each indicator, a higher value indicates a more accurate classification.

B. Comparison Methods

Compare with several advanced and representative classification methods to verify the effectiveness and superiority of DFI-Net on YRCWD. The comparison methods include a CNN-based 3-D deep learning (3DCNN) [50], a residual spectral–spatial attention network (RSSAN) [51], an attention-based bidirectional long short-term memory network (ABLSTM) [52], a transformer-based backbone network named spectral former (Speformer) [53], a group-aware hierarchical transformer (GAHT) [54], HCTNet [55], and extended vision transformer (ExViT) [56]. Considering that 3DCNN, RSSAN, ABLSTM, Speformer, and GAHT are only developed for single-source data classification, MSI and SAR are concatenated into a cube as their inputs. All of the experiments are repeated five times, and the average results with standard variations are reported. Table III lists the OA, AA, and K of each classification method on YRCWD (bold values represent the optimal classification accuracy under the same evaluation criteria).

TABLE IV
ABLATION ANALYSIS EXPERIMENT IN YRCWD

Case	Component			Indicators		
	Conv3D–2D	Cross-Agent Attention	Spe–Spa Transformer Encoder	OA(%)	AA(%)	K×100
1	-	√	√	96.02	82.23	93.69
2	√	-	√	96.86	87.34	95.02
3	√	√	-	95.99	82.82	93.65
4	√	√	√	**97.03**	**89.84**	**95.30**

The evaluation data show that the proposed DFI-Net achieves the highest OA, AA, and K in classification tasks, demonstrating superior performance. Compared with 2DCNN, 3DCNN simultaneously extracts channel–spatial features, which is more 1-D than 2DCNN. The OA value of this classification method is only 94.58%, which almost cannot classify natural willow forest, and the classification effect is poor. RSSAN designs spectral attention module and spatial attention module and embeds them into residual structure. Compared with 3DCNN, the OA improvement rate is 1.46%. The OA values of ABLSTM and Speformer are not much different from those of 3DCNN, which are 94.78% and 94.97%, respectively. GAHT uses the GAN model to show good classification performance and classification accuracy, with OA up to 96.53%. Apart from GAHT, the OAs of the MSI and SAR joint classification models HCTNet, ExViT, and DFI-Net are all higher than those of 3DCNN, RSSAN, ABLSTM, and Speformer, which use single-source data as input. This indicates that the model with branch architecture is more suitable for MSI and SAR joint classification tasks. Furthermore, DFI-Net achieves an OA value of 97.03%, which is 2.45% higher than that of 3DCNN (i.e., 94.58%), 0.99% higher than that of RSSAN (i.e., 96.04%), 2.25% higher than that of ABLSTM (i.e., 94.78%), 2.06% higher than that of Speformer (i.e., 94.97%), 0.5% higher than that of GAHT (i.e., 96.53%), 0.54% higher than that of HCTNet (i.e., 96.49%), and 0.69% higher than that of ExViT (i.e., 96.34%). It can be seen that the proposed DFI-Net exhibits the most competitive classification accuracy. In particular, DFI-Net takes the lead in the classification accuracy of various categories, which is confirmed by the calculation results of AA and K.

For the sake of intuitive analysis, the final classification diagram of all methods on the Yellow River coastal wetland dataset is shown in Fig. 6. Due to the large size of the scene, select a representative area to zoom in to investigate and highlight the details of the classification results of different methods. Compared with other methods, DFI-Net provides the classification map closest to the GT map. The visualization results of DFI-Net classification maps have the characteristics of less misclassification, low noise, accurate edges, and smooth effects. The reason why the proposed method achieves superior classification results is due to the complementary advantages of MSI and SAR features and the maximum utilization of spatial–channel information in interactive fusion features.

C. Ablation Study

Since the proposed network benefits from multiple components, we analyze the necessity of each component and its contribution to classification accuracy through a series of ablation experiments. In detail, the proposed network is mainly divided into three parts: 3D–2DCNN for feature extraction, FIFB for feature interaction fusion, and CSTE for mining deep channel–spatial information.

As we can see in Table IV, DFI-Net outperforms all its variants (bold values represent the optimal classification accuracy under the same evaluation criteria). In Case 1, the 3D–2DCNN component was removed and the original MSI and SAR data were taken directly as inputs. Compared with Case 4, the OA value of this variant decreased by 1.01%, which verified the significant effect of the convolutional module in the initial stage of feature extraction. This indicates that the spectral–spatial features extracted by 3D–2DCNN from the raw data are effective. Case 2 removed FIFB, and OA was 0.17% lower than the optimal value. Compared with Case 4, although OA decreased only slightly, AA value significantly reduced by 2.52%. This suggests that feature association is promoted in interactive fusion, and the classification accuracy of various ground objects is improved. Case 3 achieved 95.99% OA without the CSTE component, which is 1.04% lower than Case 4. This shows that DFI-Net benefits from CSTE and improves the accuracy of the model under the condition of deep information of interactive fusion features. The ablation analysis further confirms the effectiveness of the proposed network framework, showing that all of its major components are important and indispensable. Through this combination, DFI-Net achieves state-of-the-art performance in the joint classification task of MSI and SAR data.

V. DISCUSSION

A. Sensitivity Analysis of Different Patch Sizes

In practical processing, patch size determines how many surrounding pixels are used to classify center pixels, which is a basic parameter that affects classification performance. Although smaller patches can capture fine-grained details, they require additional contextual information to accurately classify. In contrast, larger patches contain more contextual information but require more computing resources. It is crucial to conduct experiments to determine the optimal patch size, which can ensure computational efficiency and improve model classification performance. So, we conduct a series of experiments. Change the patch size of the input image from small to large and observe its impact on classification accuracy. We test several different patch sizes $\{3 \times 3, 5 \times 5, 7 \times 7, 9 \times 9, 11 \times 11\}$, and the OA and time changes are shown in Fig. 7.

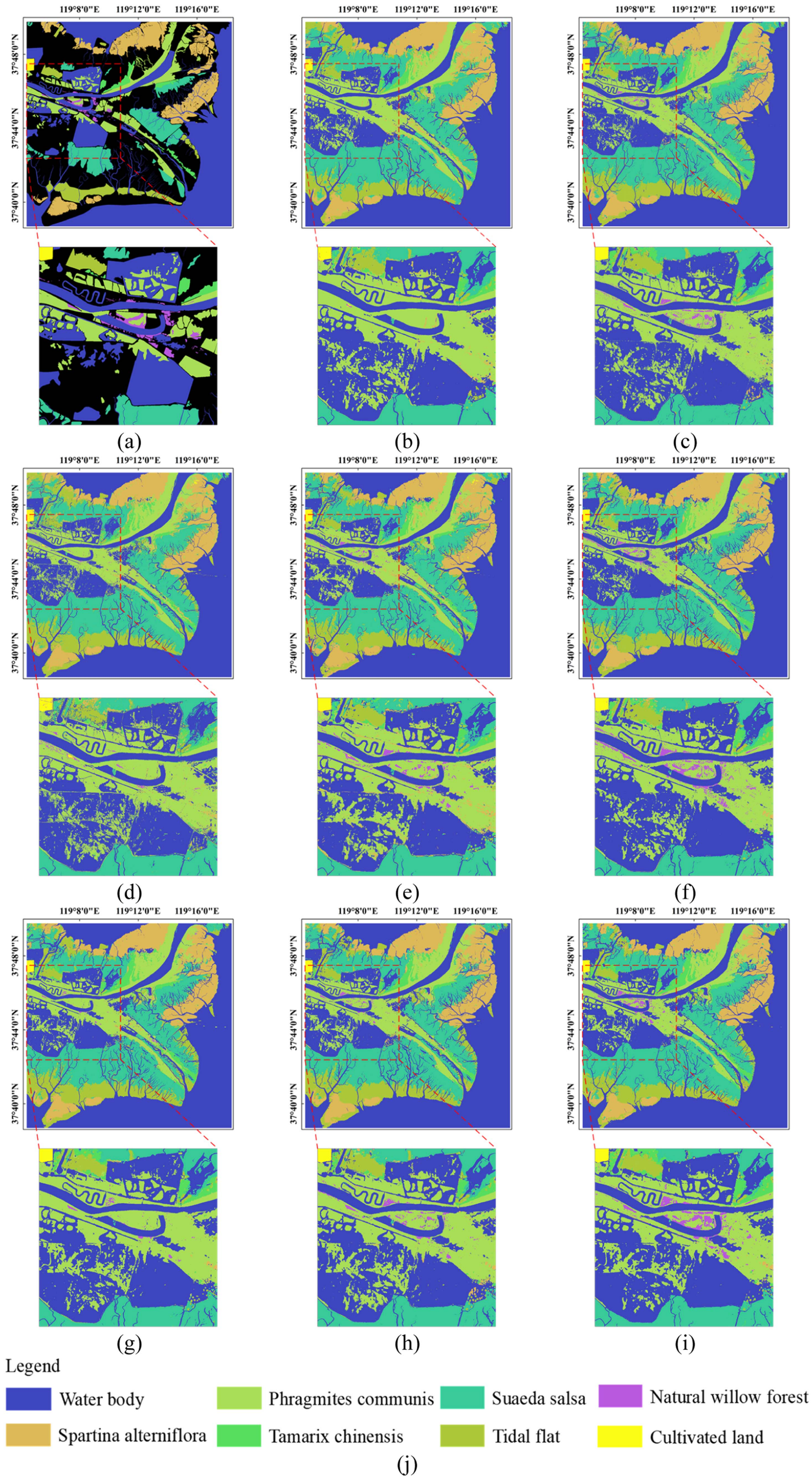


Fig. 6. (a) GT. (b) 3DCNN. (c) RSSAN. (d) ABLSTM. (e) Speformer. (f) GAHT. (g) HCTNet. (h) ExViT. (i) Proposed. (j) Legend.

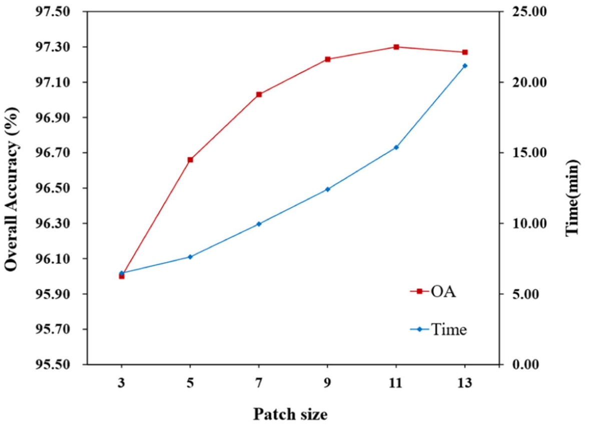


Fig. 7. Influence of different patch sizes on the OA and time.

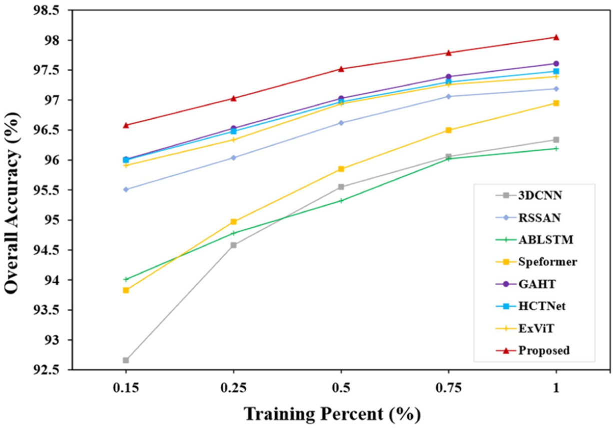


Fig. 8. OA of different methods with different numbers of training samples per class.

Fig. 7 shows the impact of different patch sizes on the OA and processing time on YRCWD. The results show that, as the patch size increases from 3 to 9, the OA steadily increases from 96.00% to 97.23%. Increase the patch size to 11, the accuracy is not significantly improved. Furthermore, increase to 13, OA shows a downward trend. During the whole process, the processing time increases from 6.48 min to approximately 21.17 min, showing a linear growth trend. Further analyzing the patch size increase from 7 to 9, the 0.2% increase in accuracy pays a time cost of up to 3 min. Therefore, compared with the 9 × 9 patch, the 7 × 7 patch is better.

The reasons for the first increase and then decrease of OA are analyzed. When the patch size is small, the model may find it challenging to capture the context information of the image. As a result, the model may be unable to fully comprehend the image features, leading to reduced classification accuracy. With the gradual increase of patch size of the input image, the model can obtain more context information and image details, thus improving the classification accuracy. However, when the patch size is too large, the redundant information in the patch can affect the discriminability of the central pixel, thereby reducing experimental performance. As for training time, a slow increase followed by a rapid increase is observed. Patch size that is too large can lead to problems of dimensionality explosion and insufficient computational resources for the model, which make both training and inference difficult and, thereby, increase training time.

In summary, the change in patch size significantly affects the classification accuracy. Appropriately increasing the patch size can improve the classification performance, but a patch size that is too large will lead to information redundancy and calculation burden. Therefore, in order to balance model accuracy and calculation overhead, we set the patch size to 7 × 7.

B. Robustness Analysis of Different Numbers of Training Samples

In wetland classification, increasing the number of training samples cannot only reduce the data imbalance but also improve the learning and generalization ability of the model. However, more training samples lead to longer training times. Therefore, achieving a balance between performance and training time with an appropriate sample size is important for optimizing wetland classification models. To verify the stability and generalization ability of the proposed method, we compare the classification performance of all methods under different training sample ratios. The 0.15%, 0.25%, 0.5%, 0.75%, and 1% of each sample are randomly selected as the training data of YRCWD. Fig. 8 shows the OA results for the different classifiers.

As can be seen from Fig. 8, the more training samples there are, the higher the classification accuracy of the model. The classification performance of all models is positively correlated with the number of training samples. Compared with several reference methods, the proposed DFI-Net maintains the best classification performance in the whole sample size range. Even in the case of limited training data, it still has the highest classification accuracy. The OA of GAHT and HCTNet is slightly lower than that of DFI-Net. ExViT is close to the OA of GAHT and HCTNet at higher training ratios (0.75% and 1%). ABLSTM and 3DCNN performed poorly in OA compared with other methods, especially 3DCNN performs the worst at low training ratios. Most models improve significantly when the training ratio goes from 0.15% to 0.25%, which we consider 0.25% as the critical training ratio.

VI. Conclusion

In this article, a dual-branch network based on feature interaction is designed for the joint classification of coastal wetlands using MSI and SAR data. FIFB realizes feature interaction and fusion from different data, which weakens the feature differences between MSI and SAR and promotes information balance. The CSTE module fully mines channel–spatial features in the high-dimensional space after interactive fusion. It enhances the expression ability of MSI and SAR fusion features and improves the overall classification accuracy. A series of experiments are conducted on the YRCWD. DFI-Net has better classification performance than the comparison model. For future work, the domain adaptive method is considered to be introduced into the model to realize cross-scene classification tasks for coastal wetlands. It can solve difficulties, such as the high cost of collecting labeled samples and inconsistent collection conditions.

ACKNOWLEDGMENT

The authors would like to thank the European Space Agency (ESA) (Copernicus Data Space Ecosystem | Europe's eyes on Earth) and the National Platform for Common GeoSpatial Information Services (Map World) for providing the data.

REFERENCES

[1] K. Liu et al., "Mapping coastal wetlands using transformer in transformer deep network on China ZY1-02D hyperspectral satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3891–3903, May 2022.

[2] Y. Liu et al., "Tracking changes in coastal land cover in the Yellow Sea, East Asia, using Sentinel-1 and Sentinel-2 time-series images and Google Earth Engine," *ISPRS J. Photogramm. Remote Sens.*, vol. 196, pp. 429–444, Feb. 2023.

[3] X. Wang et al., "Mapping coastal wetlands of China using time series Landsat images in 2018 and Google Earth Engine," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 312–326, May 2020.

[4] S. Mahdavi, B. Salehi, J. Granger, M. Amani, B. Brisco, and W. Huang, "Remote sensing for wetland classification: A comprehensive review," *GISci. Remote Sens.*, vol. 55, no. 5, pp. 623–658, Sep. 2018.

[5] F. Guo et al., "Semi-supervised cross-domain feature fusion classification network for coastal wetland classification with hyperspectral and LiDAR data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 120, Jun. 2023, Art. no. 103354.

[6] F. Guo et al., "Multisource feature embedding and interaction fusion network for coastal wetland classification with hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 2024, Art. no. 5509516.

[7] S. Gao, B.-H. Tang, L. Huang, and G. Chen, "Identification of tea plantations in typical plateau areas with the combination of Sentinel-1/2 optical and radar remote sensing data based on feature selection algorithm," *Int. J. Remote Sens.*, pp. 1–21, Apr. 2023.

[8] M. Rezaee, M. Mahdianpari, Y. Zhang, and B. Salehi, "Deep convolutional neural network for complex wetland classification using optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3030–3039, Sep. 2018.

[9] M. Amani et al., "Wetland change analysis in Alberta, Canada using four decades of Landsat imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10314–10335, Aug. 2021.

[10] Y. Yuan et al., "Multi-resolution collaborative fusion of SAR, multispectral and hyperspectral images for coastal wetlands mapping," *Remote Sens.*, vol. 14, no. 14, Jul. 2022, Art. no. 3492.

[11] Q. Feng et al., "Integrating multitemporal Sentinel-1/2 data for coastal land cover classification using a multibranch convolutional neural network: A case of the Yellow River delta," *Remote Sens.*, vol. 11, no. 9, Apr. 2019, Art. no. 1006.

[12] Y. Cai, X. Li, M. Zhang, and H. Lin, "Mapping wetland using the object-based stacked generalization method based on multi-temporal optical and SAR data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 92, Oct. 2020, Art. no. 102164.

[13] N. Liu, Q. Zhao, R. Williams, and B. Barrett, "Enhanced crop classification through integrated optical and SAR data: A deep learning approach for multi-source image fusion," *Int. J. Remote Sens.*, pp. 1–29, Jul. 2023.

[14] J. Liu et al., "Spatiotemporal change detection of coastal wetlands using multi-band SAR coherence and synergetic classification," *Remote Sens.*, vol. 14, no. 11, May 2022, Art. no. 2610.

[15] J. Muro, A. Strauch, E. Fitoka, M. Tompoulidou, and F. Thonfeld, "Mapping wetland dynamics with SAR-Based change detection in the cloud," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1536–1539, Oct. 2019.

[16] M. Gierszewska and T. Berezowski, "On the role of polarimetric decomposition and speckle filtering methods for C-band SAR wetland classification purposes," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2845–2860, Mar. 2022.

[17] S. Adeli et al., "Wetland monitoring using SAR data: A meta-analysis and comprehensive review," *Remote Sens.*, vol. 12, no. 14, Jul. 2020, Art. no. 2190.

[18] S. Niculescu et al., "Synergy of high-resolution radar and optical images satellite for identification and mapping of wetland macrophytes on the Danube delta," *Remote Sens.*, vol. 12, no. 14, Jul. 2020, Art. no. 2188.

[19] X. Wang et al., "Contribution of land cover classification results based on Sentinel-1 and 2 to the accreditation of wetland cities," *Remote Sens.*, vol. 15, no. 5, Feb. 2023, Art. no. 1275.

[20] M. Mahdianpari, B. Salehi, F. Mohammadimanesh, S. Homayouni, and E. Gill, "The first wetland inventory map of Newfoundland at a spatial resolution of 10 m using Sentinel-1 and Sentinel-2 data on the Google Earth Engine cloud computing platform," *Remote Sens.*, vol. 11, no. 1, Dec. 2018, Art. no. 43.

[21] B. Slagter, N. E. Tsendbazar, A. Vollrath, and J. Reiche, "Mapping wetland characteristics using temporally dense Sentinel-1 and Sentinel-2 data: A case study in the St. Lucia wetlands, South Africa," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 86, Apr. 2020, Art. no. 102009.

[22] M. Wang et al., "Wetland mapping in East Asia by two-stage object-based random forest and hierarchical decision tree algorithms on Sentinel-1/2 images," *Remote Sens. Environ.*, vol. 297, Nov. 2023, Art. no. 113793.

[23] T. Berhane et al., "Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory," *Remote Sens.*, vol. 10, no. 4, Apr. 2018, Art. no. 580.

[24] X. Wang et al., "Land-cover classification of coastal wetlands using the RF algorithm for Worldview-2 and Landsat 8 images," *Remote Sens.*, vol. 11, no. 16, Aug. 2019, Art. no. 1927.

[25] A. Jamali, M. Mahdianpari, B. Brisco, J. Granger, F. Mohammadimanesh, and B. Salehi, "Comparing solo versus ensemble convolutional neural networks for wetland classification using multi-spectral satellite imagery," *Remote Sens.*, vol. 13, no. 11, May 2021, Art. no. 2046.

[26] B. Fu et al., "Quantifying scattering characteristics of mangrove species from optuna-based optimal machine learning classification using multi-scale feature selection and SAR image time series," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, Aug. 2023, Art. no. 103446.

[27] H. Xing, J. Niu, Y. Feng, D. Hou, Y. Wang, and Z. Wang, "A coastal wetlands mapping approach of Yellow River Delta with a hierarchical classification and optimal feature selection framework," *CATENA*, vol. 223, Apr. 2023, Art. no. 106897.

[28] B. Fu et al., "Mangrove species classification using novel adaptive ensemble learning with multi-spatial-resolution multispectral and full-polarization SAR images," *Int. J. Digit. Earth*, vol. 17, no. 1, Dec. 2024, Art. no. 2346277.

[29] B. Fu et al., "Quantifying vegetation species functional traits along hydrologic gradients in karst wetland based on 3D mapping with UAV hyperspectral point cloud," *Remote Sens. Environ.*, vol. 307, Jun. 2024, Art. no. 114160.

[30] B. Fu et al., "Combination of super-resolution reconstruction and SGA-net for marsh vegetation mapping using multi-resolution multispectral and hyperspectral images," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 2724–2761, Jul. 2024.

[31] J. V. Solórzano, J. F. Mas, Y. Gao, and J. A. Gallardo-Cruz, "Land use land cover classification with U-Net: Advantages of combining Sentinel-1 and Sentinel-2 imagery," *Remote Sens.*, vol. 13, no. 18, Sep. 2021, Art. no. 3600.

[32] G. Konapala, S. V. Kumar, and S. K. Ahmad, "Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 163–173, Oct. 2021.

[33] D. Ienco, R. Interdonato, R. Gaetano, and D. Ho Tong Minh, "Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 11–22, Dec. 2019.

[34] 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.

[35] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and SAR image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 8057–8070, Oct. 2023.

[36] E. R. DeLancey, J. F. Simms, M. Mahdianpari, B. Brisco, C. Mahoney, and J. Kariyeva, "Comparing deep learning and shallow learning for large-scale wetland classification in Alberta, Canada," *Remote Sens.*, vol. 12, no. 1, Dec. 2019, Art. no. 2.

[37] A. Jamali, M. Mahdianpari, F. Mohammadimanesh, B. Brisco, and B. Salehi, "A synergic use of Sentinel-1 and Sentinel-2 imagery for complex wetland classification using generative adversarial network (GAN) scheme," *Water*, vol. 13, no. 24, Dec. 2021, Art. no. 3601.

[38] A. Jamali, M. Mahdianpari, B. Brisco, D. Mao, B. Salehi, and F. Mohammadimanesh, "3DUNetGSFormer: A deep learning pipeline for complex wetland mapping using generative adversarial networks and Swin transformer," *Ecol. Inf.*, vol. 72, Dec. 2022, Art. no. 101904.

[39] A. Jamali, S. K. Roy, and P. Ghamisi, "WetMapFormer: A unified deep CNN and vision transformer for complex wetland mapping," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 120, Jun. 2023, Art. no. 103333.

[40] Y. Gao et al., "Fusion classification of HSI and MSI using a spatial-spectral vision transformer for wetland biodiversity estimation," *Remote Sens.*, vol. 14, no. 4, Jan. 2022, Art. no. 850.

[41] X. Zhao et al., "Fractional Fourier image transformer for multimodal remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2314–2326, Feb. 2024.

[42] B. Hosseiny, M. Mahdianpari, B. Brisco, F. Mohammadimanesh, and B. Salehi, "WetNet: A spatial–temporal ensemble deep learning model for wetland classification using Sentinel-1 and Sentinel-2," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 4406014.

[43] A. Jamali and M. Mahdianpari, "Swin transformer and deep convolutional neural networks for coastal wetland classification using Sentinel-1, Sentinel-2, and LiDAR data," *Remote Sens.*, vol. 14, no. 2, Jan. 2022, Art. no. 359.

[44] H. Jafarzadeh, M. Mahdianpari, and E. W. Gill, "Wet-GC: A novel multimodel graph convolutional approach for wetland classification using Sentinel-1 and 2 imagery with limited training samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5303–5316, Jun. 2022.

[45] H. Lin, X. Cheng, X. Wu, and D. Shen, "CAT: Cross attention in vision transformer," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2022, pp. 1–6.

[46] D. Han et al., "Agent attention: On the integration of softmax and linear attention," Dec. 2023, *arXiv:2312.08874*.

[47] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.

[48] R. Zhang et al., "A comparison of Gaofen-2 and Sentinel-2 imagery for mapping mangrove forests using object-oriented analysis and random forest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4185–4193, Apr. 2021.

[49] A. Jamali, M. Mahdianpari, F. Mohammadimanesh, and S. Homayouni, "A deep learning framework based on generative adversarial networks and vision transformer for complex wetland classification using limited training samples," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, Dec. 2022, Art. no. 103095.

[50] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

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

[52] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5509612.

[53] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5518615.

[54] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5539014.

[55] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and LiDAR data using a hierarchical CNN and transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Dec. 2023, Art. no. 5500716.

[56] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 5514415.

Mingming Xu (Member, IEEE) received the B.S. degree in surveying and mapping engineering from the China University of Petroleum (East China), Qingdao, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2016.

She is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China). Her research interests include hyperspectral image processing and wetland remote sensing.

Mingwei Liu received the B.S. degree in engineering from the Inner Mongolia University of Science and Technology, Baotou, China, in 2022. She is currently working toward the master's degree in surveying and mapping science and technology with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China.

Her research direction is remote sensing of wetlands.

Yanfen Liu received the B.S. degree in surveying and mapping engineering from the University of Petroleum (East China), Dongying, China, in 2003, and the Ph.D. degree in environmental science from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2010.

She is currently a Senior Engineer with Dongying Marine Development Research Institute, Dongying, China. Her research interests include wetland remote sensing, nature reserves, and marine biodiversity.

Shanwei Liu received the Ph.D. degree in environmental science from the Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China. His research interests include satellite altimetry, ocean remote sensing, hyperspectral image processing, and GIS application.

Hui Sheng received the Ph.D. degree in geological resources and geological engineering from the China University of Petroleum, Qingdao, China, in 2010.

He is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China). His research interests include ocean remote sensing, hyperspectral image processing, and photogrammetry.