Mapping Coastal Wetlands Using Transformer iN Transformer Deep Network on China ZY1-02D Hyperspectral Satellite Images

Kai Liu, Weiwei Sun, *Senior Member, IEEE*, Yijun Shao, Weiwei Liu, Gang Yang, *Member, IEEE*, Xiangchao Meng, *Member, IEEE*, Jiangtao Peng, *Senior Member, IEEE*, Dehua Mao, and Kai Ren

***Abstract*—Coastal wetlands mapping is a big challenge in remote sensing fields because of similar spectrum of different ground objects and their severe fragmentation and spatial heterogeneity. In this article, we propose a hyperspectral image transformer iN transformer (HSI-TNT) method for mapping coastal wetlands on ZiYuan1-02D (ZY1-02D) hyperspectral images, which uses two transformer deep networks to fuse local and global features. First, we put forward the idea that each hyperspectral pixel can be considered as a superpixel in spectral dimension, and subsequent position encodings are employed aiming to retain spatial information. After that, in each HSI-TNT block, the local information between pixels is extracted by inner T-Block, and added to the patch space by linear transformation to extract the global information by outer T-Block. Finally, the stacked HSI-TNT block, also known as HSI-TNT framework, is used for classification and mapping. Experimental results show that HSI-TNT achieves the best results on both Yancheng and Yellow River Delta wetlands data, with overall classification accuracy of 95.57% and 93.69%, respectively. The HSI-TNT combined with ZY1-02D satellite hyperspectral data has huge potentials in mapping coastal wetlands.**

***Index Terms*—Classification, coastal wetlands, hyperspectral image transformer in transformer, hyperspectral remote sensing, ZY1-02D satellite.**

Manuscript received March 10, 2022; revised April 15, 2022; accepted April 28, 2022. Date of publication May 10, 2022; date of current version May 25, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42122009, Grant 41971296, Grant 41801252, Grant 41801256, and Grant 61871177, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR19D010001 and Grant LQ18D010001, in part by the Natural Science Foundation of Hubei Province under Grant 2021CFA087, in part by the Public Science and Technology Plan Projects of Ningbo City under Grant 2021S089, in part by the Science and Technology Innovation 2025 Major Project of Ningbo City under Grant 2021Z107 and Grant 2022Z032, and in part by the Zhejiang Province General Research Project of China under Grant Y202148270. *(Corresponding authors: Weiwei Sun; Yijun Shao; Weiwei liu.)*

Kai Liu, Weiwei Sun, Weiwei Liu, Gang Yang, and Kai Ren are with the Department of Geography and Spatial Information Techniques, Ningbo University, Ningbo 315211, China (e-mail: liukai1726867269@163.com; nbsww@outlook.com; liuweiwei@nbu.edu.cn; yanggang@nbu.edu.cn; 15158346549@163.com).

Yijun Shao is with the School of Material Science and Chemical Engineering, Ningbo University, Ningbo 315211, China (e-mail: shaoyijun@nbu.edu.cn).

Xiangchao Meng is with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China (e-mail: mengxiangchao@nbu.edu.cn).

Jiangtao Peng is with the Hubei Key Laboratory of Applied Mathematics, Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China (e-mail: pengjt1982@hubu.edu.cn).

Dehua Mao is with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China (e-mail: maodehua@iga.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3173349

I. Introduction

COASTAL wetlands are transition zones between terrestrial and marine ecosystems [1], consisting of plants, animals, microorganisms, and associated environments. They have great ecological and economic value in regulating runoff, nitrogen fixation, preventing seawater intrusion, and supplementing groundwater [2]–[4]. However, coastal wetlands are facing serious threats such as sea level rise [5], land use conversion [6], and invasion by alien plants under the condition of accelerated global warming and population growth [7], which put many coastal wetlands at risks of degradation or even disappearance [8], [9]. Therefore, it is of great significance for mapping coastal wetlands and for further resource utilization and ecological protection.

Remote sensing with the advantages of high temporal and spatial resolution, could effectively eliminate the limitation of time and labor consuming of *in situ* investigation [10]. Panchromatic image has been extensively used in ground objects observation, but it has only a single spectral channel, which restricts the accurate detection of ground objects. Multispectral image only contains several discrete spectral channels in the visible to near-infrared [11], [12], making it impossible to distinguish objects with highly similar spectrum (e.g., Suaeda and reed). In contrast, hyperspectral image (HSI) contains hundreds of continuous spectral channels, which has both rich spatial information and spectral features of ground objects, and is widely used in fields such as classification and oil spill detection [13]–[16]. Accordingly, applying hyperspectral remote sensing has a greater potential to identify complex coastal wetlands objects.

Currently, supervised classification is the most commonly used method for mapping coastal wetlands by HSI, which is to classify unknown pixels by establishing a discriminant function from the labeled samples [17]. For spectral methods, features are extracted (e.g., by PCA) [18] or selected (e.g., by Bhattacharyya distance) [19] and then passed to classifiers (e.g., SVM) [20] for classification. For spatial methods, extended morphological profiles [21] and gray level co-occurrence matrix [22] implement

spatial filters to extract the spatial dependence of ground objects. Although a dual spatial information fusion HSI classification framework is proposed to improve the classification results in [23], these methods only consider the spectral or spatial information separately while ignoring their global spatial-spectral structure. In contrast, three-dimensional (3-D) wavelet filters [24] and 3-D Gabor filters [25] utilize the spatial-spectral features to improve classification accuracy. Unfortunately, they are constrained by the underlying or middle layer features of HSI data and have scant feature characterization capability. That accordingly bring about "semantic gap" especially for the complicated coastal wetlands environment. Moreover, they are only applicable to certain specific scenarios and have poor generalization capability.

Recently, deep learning models have also been applied to HSI classification, such as deep belief network (DBN) [26], stacked autoencoder (SAE) [27], and convolutional neural networks (CNNs) [28], [29]. Nevertheless, DBN and SAE require to flatten the input image block into a 1-D vector, which would destroy spatial structure of original images [30]. CNNs have the characteristics of sparse connection, parameter sharing and equivariant mapping, which reduces the network complexity and the training parameter sizes, and can extract spatial-spectral features through sliding convolution [31]. Double-branch dual-attention mechanism network uses a unified convolution to extract the spectral-spatial information directly from the original HSI, avoiding the loss of information [32]. In [16], a multilayer global spatial-spectral attention network based on UAV-hyperspectral dataset is proposed for coastal wetlands mapping and achieves the optimal performance. In addition, multisource remote sensing data as the input of deep learning method can further improve the accuracy of land cover classification, such as asymmetric feature fusion network based on hyperspectral and SAR [33], depthwise feature interaction network based on hyperspectral and multispectral [34], and interleaving perception convolutional neural network (IP-CNN) based on hyperspectral and LiDAR [35]. These models all show remarkable performance in HSI classification.

Despite the great success of CNN-based models in HSI, the implementation in coastal wetlands mapping still suffers from the following serious problems. First, CNNs use convolutional kernels to extract high-level features, but the convolutional kernels are limited by the local receptive fields, which makes it difficult to capture sequence information, especially middle- and long-term dependencies, which would increase the difficulty of extracting features in complex coastal wetlands. Although deeper convolutional layers can be superimposed for feature extraction, e.g., VGG16 [36], that would make the model more complex and computationally expensive. Moreover, CNNs have good spatial information extraction capability, but since the mixed components of coastal wetlands, they are inevitably susceptible to the influence of adjacent pixels when convolutionally extracting local features, which will largely limit the performance of HSI image classification.

The transformer network can solve the above-mentioned problems well [37], and has been initially applied in HSI classification, such as HSI classification bidirectional encoder representation from transformers [38], spatial-spectral transformer [39], and spatial-spectral vision transformer [40] models. These successful application of the transformer models in HSI classification benefits from the parallelization between modules and the self-attention mechanism. Parallel operation increases the efficiency of model training and conforms to the modern distributed GPU framework; the self-attention mechanism connects the distance between any two positions of the given data and retains long-distance information. Satellite hyperspectral data provides large-scale wetlands images with over 100 continuous spectral bands, but a patch usually has 7×7 or even more pixels, so multiple ground objects are commonly present in a patch. The local and structural information of hyperspectral data is particularly important for mapping coastal wetlands. However, the abovementioned models need to project the patch into a vector, and hence, local spatial structure information is corrupted, which limits the classification performance in large-scale complicated wetlands areas.

In this article, we propose the HSI transformer iN transformer (HSI-TNT) framework that fuses local and global features for mapping coastal wetlands with hyperspectral data. Our idea is to split the input patch into mini-patch sequences, each mini-patch is in turn reshaped to superpixel sequences, and then position encodings (PE) are added to preserve spatial information. The method contains two transformer blocks, named inner T-block and outer T-block, for extracting local features of the superpixel sequences and global features of the mini-patch sequences, and the local features are then added to the global features by linear projection, which can increase the local information of the input patch. In inner T-Block, all channels of an HSI pixel are regarded as a superpixel sequence, which effectively avoids the influence of the adjacent pixel interference and help to extract spectral sequence features. Our main contributions can be summarized as following.

1) We propose an innovative HSI-TNT classification framework based on structural nesting. Outer T-block models the relationship between mini-patch and inner T-block models the relationship between pixels. After linear transformation, the pixel-level features are projected into the space where the mini-patch is located and added them together to avoid the local information lost in regular transformer-based HSI classification.
2) The HSI-TNT framework uses position encodings to preserve the position information of the input data, which solves the problem of the irrelevant order of CNN-based models, and avoid the spatial feature loss in the network. Moreover, pixel-by-pixel unfold degrades the negative influence of surrounding pixels on target pixels because of spectral divergences.
3) Experimental results on Chinese hyperspectral data demonstrate that the HSI-TNT is easy to parameterize, robust, and has good generalization capabilities.

II. Methodology

A. Overview

In this article, we propose an HSI classification framework named HSI-TNT, as shown in Fig. 1(a). First, we use unfold

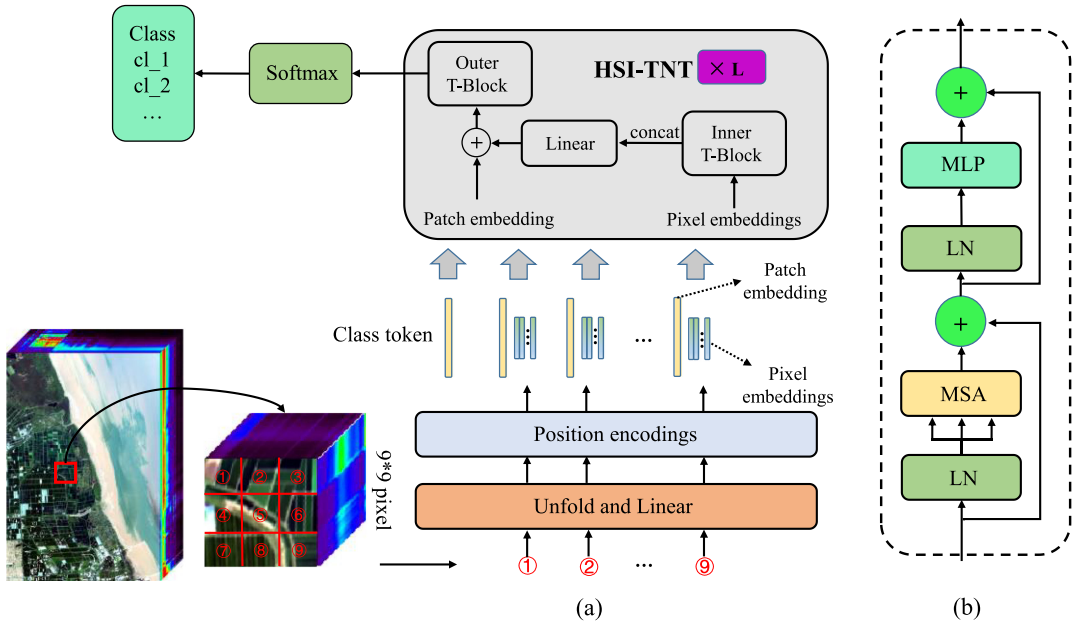


Fig. 1. Flowchart of our HSI-TNT method. (a) HSI-TNT framework. (b) Outer(Inner) T-Block.

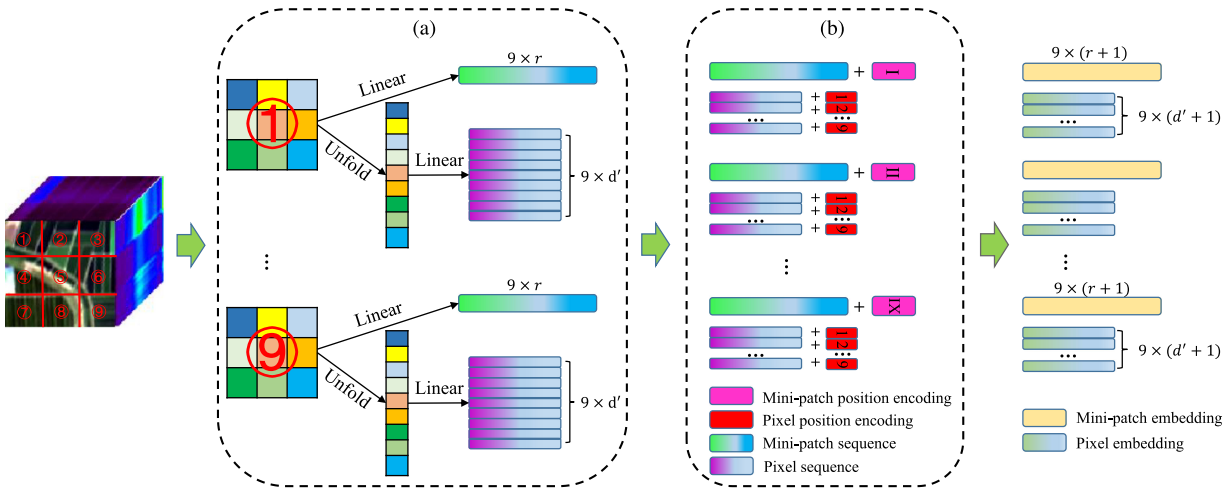


Fig. 2. (a) Unfold and linear. (b) Mini-patch and pixel sequence position encodings.

and linear to obtain mini-patch sequences and pixel sequences from a patch, which can retain the spectral information of the pixels. After that, PE is implemented on them to get mini-patch embedding and pixel embeddings to preserve spatial position information. Third, pixel embeddings are input to the inner T-Block of HSI-TNT to extract local information, which is added to the mini-patch embedding by linear transformation, in order to extract global information by outer T-Block, where outer (inner) T-Block is the transformer shown in Fig. 1(b). Finally, the class token is classified by softmax.

B. HSI-TNT Framework

In this section, the process of HSI-TNT is first described in detail, as shown in Fig. 1(a). Then, we introduce the basic

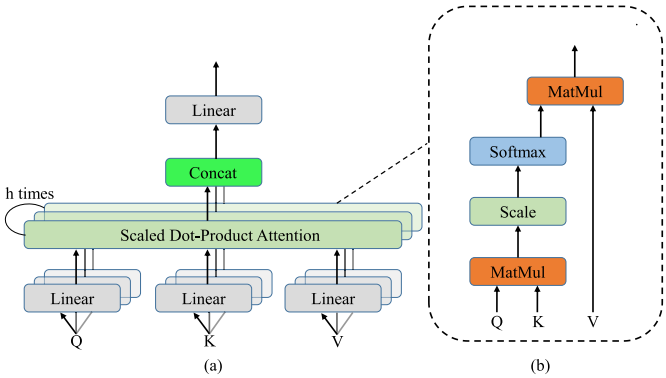


Fig. 3. MSA algorithm flowchart. (a) Multi-head Self-Attention. (b) Scaled Dot-Product Attention.

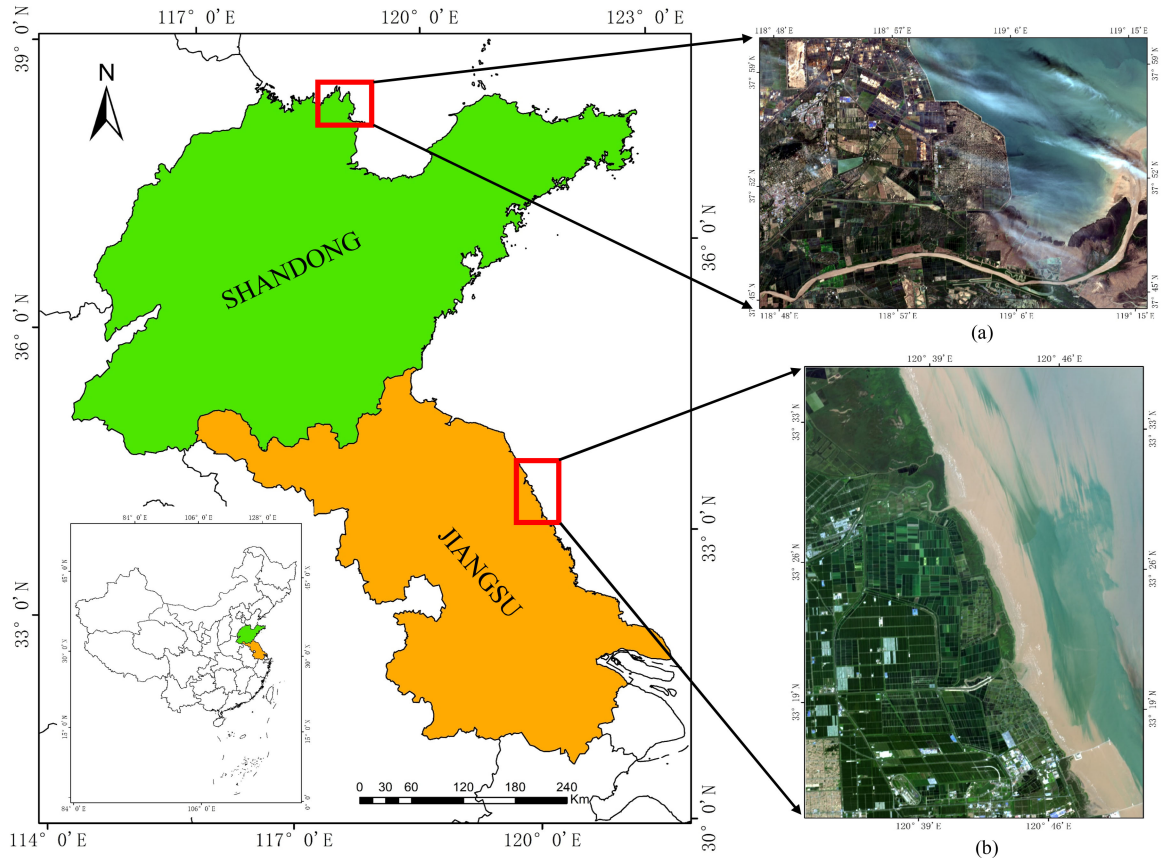


Fig. 4. Location of YRD and YC coastal wetlands. (a) Yellow River Delta. (b) Yancheng Coastal Wetlands.

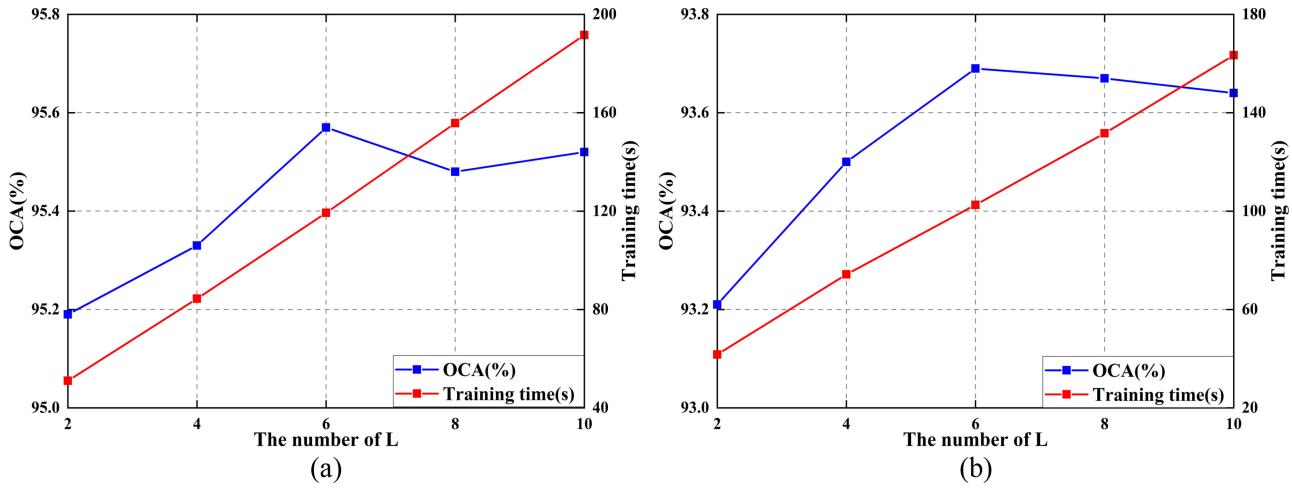


Fig. 5. Effects of L on OCA and training time for (a) YC and (b) YRD.

components of outer (inner) T-Block, including layer normalization (LN), multihead self-attention (MSA), and multilayer perceptron (MLP), as shown in Fig. 1(b). Finally, the process of unfold and linear (UL) and PE are presented in Fig. 2.

1) Hyperspectral Image Transformer iN Transformer: For a given hyperspectral data $\boldsymbol{X} \in \mathbb{R}^{p\times p\times d}$, where p and d are the spatial size and the number of bands, respectively. In this framework, each patch needs to be further subdivided into n mini-patch $\boldsymbol{X'} = [X_1', X_2', \ldots, X_n'] \in \mathbb{R}^{n\times p'\times p'\times d}$. Then, pixel-by-pixel UL projection of $\boldsymbol{X'}$ is performed to preserve spectral sequence information [see Fig. 2(a)], showing as

$$\mathcal{Y}_0 = \left[Y_0^1, Y_0^2, \ldots, Y_0^n\right] = UL\left(\boldsymbol{X'}\right) \in \mathbb{R}^{n\times m\times d'} \tag{1}$$

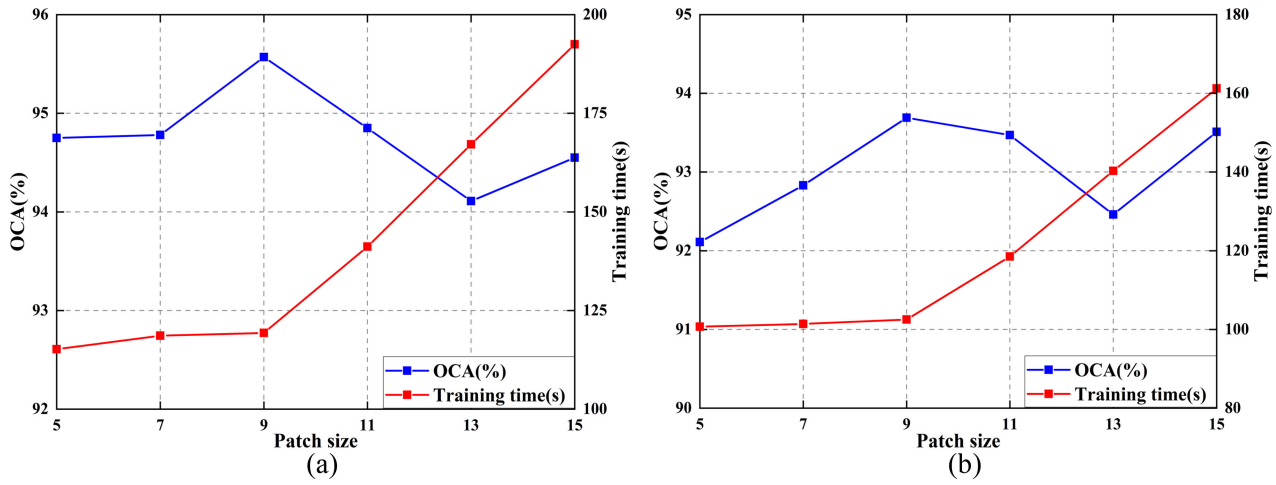


Fig. 6. Effects of different size of patch on OCA and training time for (a) YC and (b) YRD.

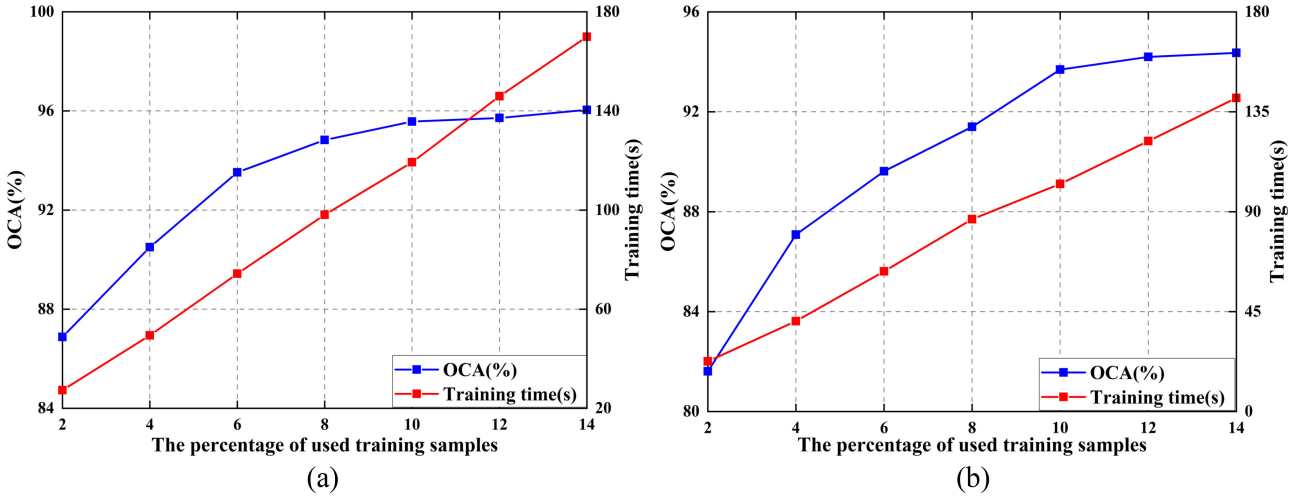


Fig. 7. Effects of different training samples percentage on OCA and training time for (a) YC and (b) YRD.

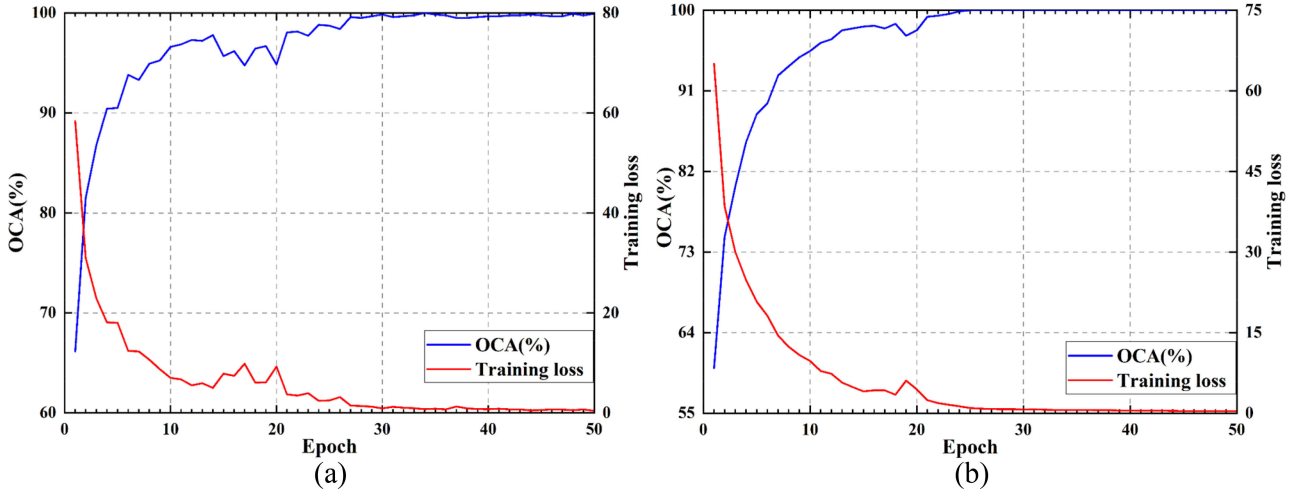


Fig. 8. Training loss curve and OCA for (a) YC and (b) YRD.

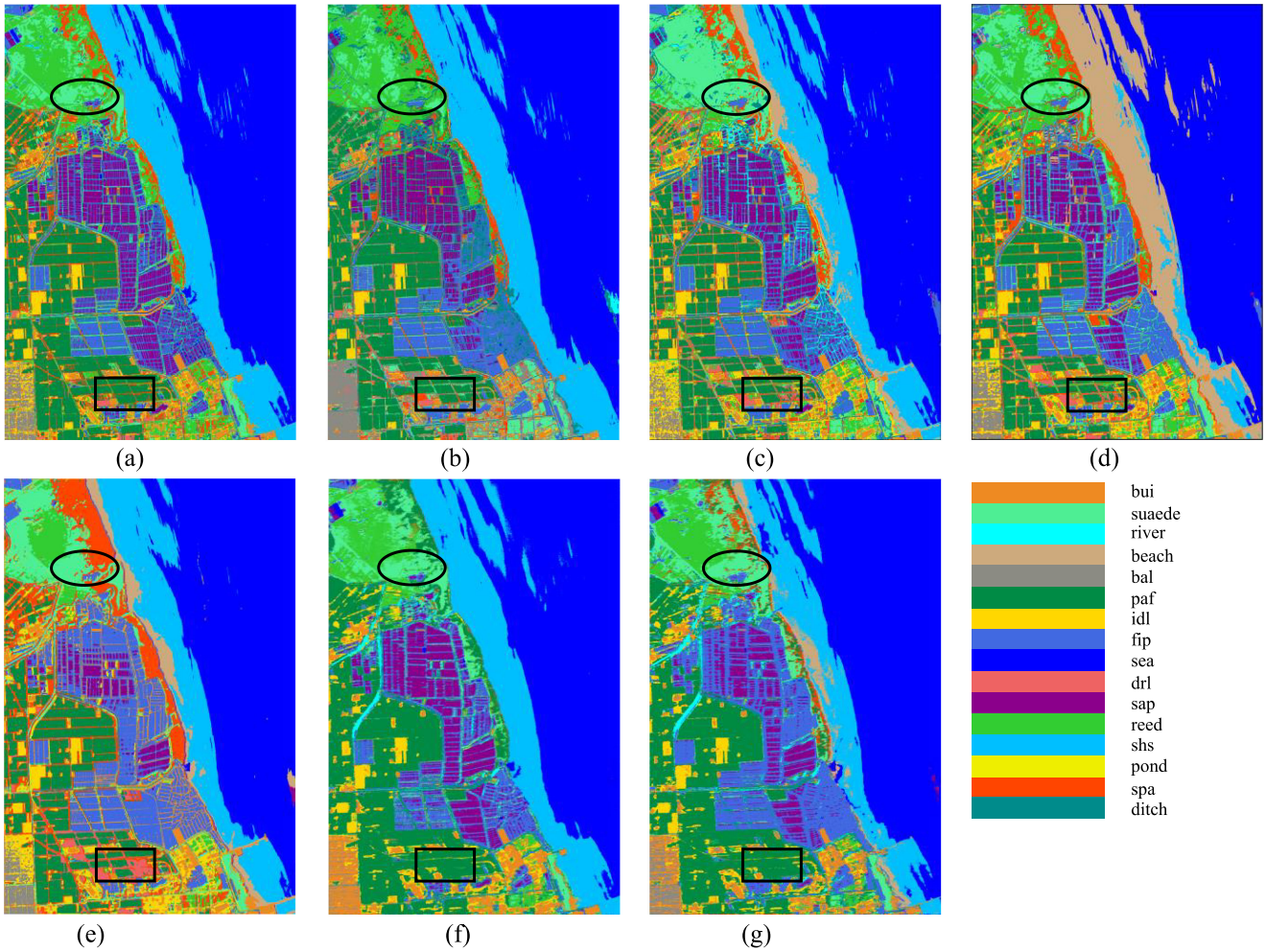


Fig. 9. Classification maps obtained by different methods of YC with the corresponding OCA. (a) SVM, OCA = 87.29%. (b) SGD, OCA = 91.14%. (c) CD-CNN,OCA = 91.79%. (d) 3D-CNN, OCA = 92.91%. (e) SSFCN-CRF, OCA = 93.81%. (f) ViT, OCA = 93.46%. (g) HSI-TNT, OCA = 95.57%.

where $Y_0^i \in \mathbb{R}^{m\times d'}, i = 1, 2, \ldots, n$, m, and $d'$are $p'^2$and the dimensions after linear projection, respectively. We regard each mini-patch Y_0^i as a sequence of pixel embeddings

$$Y_0^i = \left[y_0^{i,1}, y_0^{i,2}, \ldots, y_0^{i,m}\right]. \tag{2}$$

In HSI-TNT block, a two-step data processing is proposed for the inner T-Block and outer T-Block, defined as T_{in} and T_{out}, respectively, where outer (inner) T-Block denotes transformer block. For T_{in}, we use a transformer to study the relationship between pixels

$$Y_l'^i = Y_{l-1}^i + \text{MSA}\left(\text{LN}\left(Y_{l-1}^i\right)\right) \tag{3}$$

$$Y_l^i = Y_l' i + \text{MLP}\left(\text{LN}\left(Y_l'^i\right)\right) \tag{4}$$

where $l = 1, 2, \ldots, L$ is the lth layer, and L is the total number of layers. All Y_0^i after T_{in} are $y_l = [Y_l^1, Y_l^2, \ldots, Y_l^n]$. This procedure establishes the relationship between pixels by calculating the interaction between any two pixels.

For T_{out}, we create the mini-patch embedding memories to store the mini-patch level representation sequence: $Z_0 = [Z_{\text{class}}, Z_0^1, Z_0^2, \ldots, Z_0^n] \in \mathbb{R}^{(n+1)\times r}$, where Z_{class} (i.e., class token) is a learnable sequence embedding and they are all initialized as 0. In each layer, Y_0^i is linearly transformed and added to the embedding domain of the mini-patch level

$$Z_{l-1}^i = Z_{l-1}^i + Vec\left(Y_{l-1}^i\right) W_{l-1} + b_{l-1} \tag{5}$$

where $Vec(\cdot)$ flattens the data into a vector, b_{l-1} and W_{l-1} are the bias and weights, respectively. We use a standard mini-patch embedding transformer to establish the relationship between mini-patch embeddings

$$Z_l'^i = Z_{l-1}^i + \text{MSA}\left(\text{LN}\left(Z_{l-1}^i\right)\right) \tag{6}$$

$$Z_l^i = Z_l'^i + \text{MLP}\left(\text{LN}\left(Z_l'^i\right)\right). \tag{7}$$

To summarize, the input and output of HSI-TNT are as

$$\mathcal{Y}_l,\ Z_l = \text{HSI\_TNT}\left(\mathcal{Y}_{l-1}, Z_{l-1}\right). \tag{8}$$

In HSI-TNT, inner T-Block is used to establish the relationship between pixels to extract local features, and outer T-Block is used to establish the relationship between mini-patch to extract global features. By stacking the HSI-TNT block L times, deep features are extracted and HSI-TNT network is built. Finally,

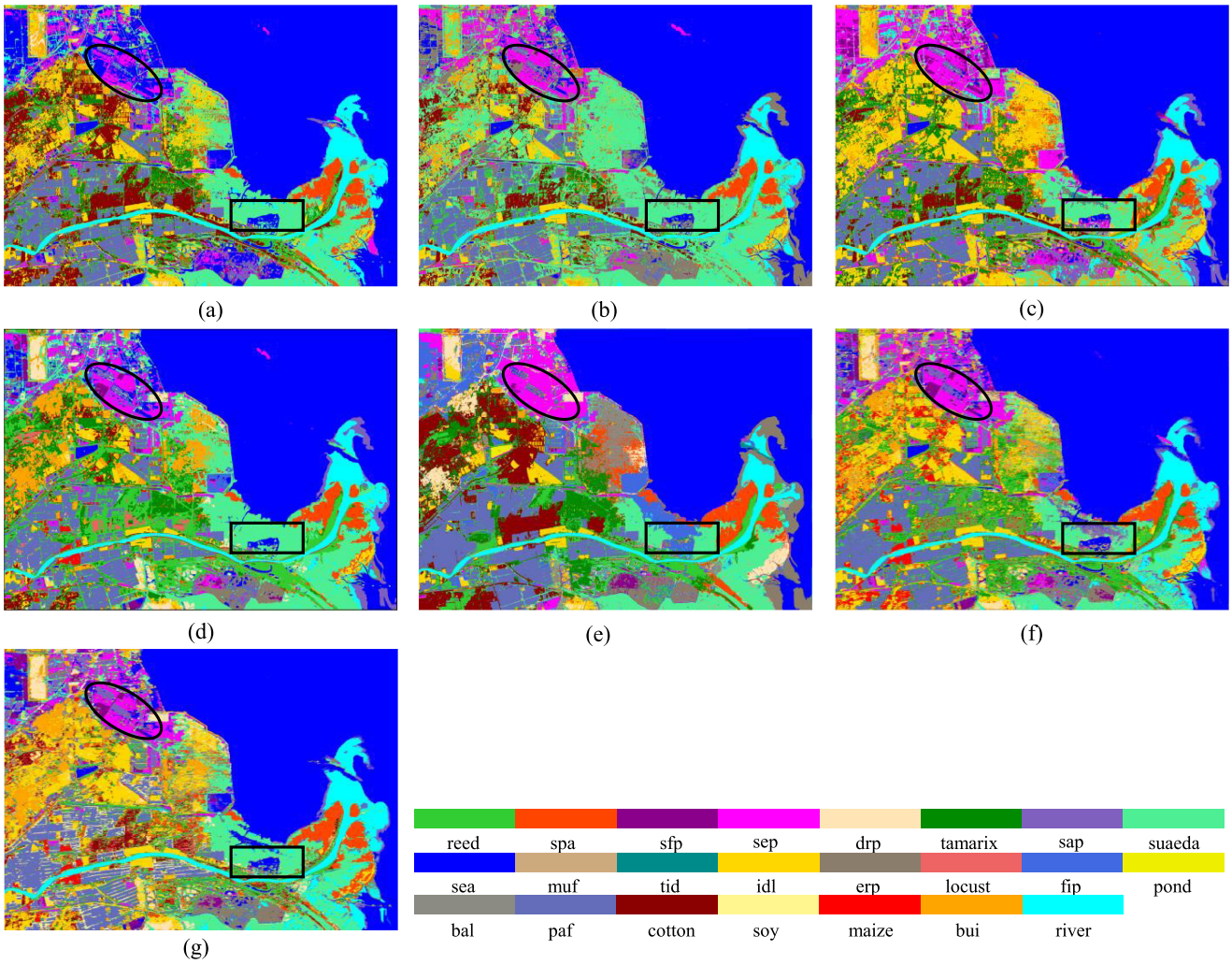


Fig. 10. Classification maps obtained by different methods of YRD with the corresponding OCA. (a) SVM, OCA = 80.97%. (b) SGD, OCA = 84.93%. (c) CD-CNN, OCA = 80.54%. (d) 3D-CNN, OCA = 90.27%. (e) SSFCN-CRF, OCA = 91.35%. (f) ViT, OCA = 90.79%. (g) HSI-TNT, OCA = 93.69%.

class token is classified as the representation of the patch by the softmax.

2) Basic Components:

1) LN: In the transformer network, the samples are normalized by LN, which not only reduces the computation time, but also alleviates the problem of vanishing or exploding gradients. It is applied to all samples $x \in \mathbb{R}^d$, showing as

$$\text{LN}\ (x) = \frac{x-\mu}{\delta} \circ \gamma + \beta \tag{9}$$

where $\delta, \mu \in \mathbb{R}$ are the standard deviation and mean of the features, respectively, $\circ$ is the elementwise dot, $\gamma, \beta \in \mathbb{R}^d$ are learnable affine transformation parameters.

2) MSA: The MSA algorithm is the central component of the transformer, which aims to capture the relevant importance of the input sequence, as shown in Fig. 3(a). Queries $Q \in \mathbb{R}^{n\times d_k}$, values $V \in \mathbb{R}^{n\times d_v}$ and keys $K \in \mathbb{R}^{n\times d_k}$ are linearly transformed from $\boldsymbol{X'} \in \mathbb{R}^{n\times d}$, where n is the length of the input sequence, d, d_v, d_k are the dimensions of the inputs, values, and queries (keys), respectively. Scaled dot-product attention [see Fig. 3(b)] is applied to combine Q, K, V as

$$\text{Attention}\,(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{10}$$

The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key [41]. Different and learned projections are used to project queries, keys, and values repeatedly (h times), and then these results are connected as given in

$$\text{MultiHead}\,(Q,K,V) = \text{concat}\,(\text{head}_1,\ldots,\text{head}_h)\,W^O \tag{11}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, $W_i^Q, W_i^K \in \mathbb{R}^{d\times d_k}$, $W_i^V \in \mathbb{R}^{d\times d_v}$, and $W^O \in \mathbb{R}^{h\times d_v\times d}$ are parameter matrix.

3) MLP: MLP is used for feature transformation and nonlinearity between self-attention layers, shown as

$$\text{MLP}\ (\boldsymbol{X'}) = \sigma\,(\boldsymbol{X'}W_1 + b_1)\,W_2 + b_2 \tag{12}$$

where $W_1 \in \mathbb{R}^{d\times d_m}$ and $W_2 \in \mathbb{R}^{d_m\times d}$ are weights of the two fully-connected layers, respectively. $b_1 \in \mathbb{R}^{d_m}$ and $b_2 \in \mathbb{R}^d$ are the bias terms, and $\sigma(\cdot)$ is the Gaussian error linear units activation function.

3) UL and PE:

1) UL: The UL operation is performed on a patch with a size of 9×9 pixels, so that T_{in} and T_{out} can extract local and global features, respectively, as shown in Fig. 2(a). Unfold will perform pixel-by-pixel operation on the mini-patch in the form of left-to-right, top-to-bottom, and project it to a vector with d' dimension through linear. That process can preserve the spectral sequence information. In addition, each pixel in mini-patch is directly projected to a vector with r dimension to store the mini-patch level sequence information. In this article, d' and r are set to 64 and 128, respectively.
2) PE: The spatial information of the input data can be preserved by adding PE, which details are shown in Fig. 2(b). UL is used to get the mini-patch sequence based on patch and pixel sequence, then corresponding standard learnable 1-D PE are added to retain the spatial information, and finally get the mini-patch embedding and pixel embeddings. It is worth noting that for the pixel sequence, the mini-patch sequence is unfolded by pixel, which is more conducive to the inner T-Block extracting spectral features. A PE is assigned to each mini-patch

$$Z_0 \leftarrow Z_0 + E_{\text{mini-patch}} \tag{13}$$

where $E_{\text{mini-patch}} \in \mathbb{R}^{(n+1)\times r}$ are mini-patch PE. For pixels in a mini-patch, pixel encodings are added to pixel embeddings

$$Y_0^i \leftarrow Y_0^i + E_{\text{pixel}}, i = 1, 2, \ldots, n \tag{14}$$

where $E_{\text{pixel}} \in \mathbb{R}^{m\times d'}$ are pixel PE. In this way, the mini-patch encodings can obtain global spatial information, while the pixel encodings are used to obtain local relative information.

III. Experimental Data and Study Area

A. ZY1-02D Hyperspectral Data

The ZiYuan1-02D (ZY1-02D) satellite, launched on September 12, 2019, is the first self-built commercial hyperspectral satellite in China [42]. It can be utilized to large-scale observation and quantitative remote sensing missions with high spectral resolution and medium spatial resolution. The advanced hyperspectral imager (AHSI), a payload of the satellite, has an imaging band of 0.4 to 2.5 μm and 166 spectral bands, including 76 spectral bands in visible near-infrared (VNIR) and 90 spectral bands in short-wave infrared (SWIR) [17]. The spatial resolution of AHSI is 30 m and the spectral resolution of VNIR bands is 10 nm while that of SWIR is 20 nm, respectively.

B. Yancheng and Yellow River Delta Coastal Wetlands

The Yancheng (YC) and Yellow River Delta (YRD) in Fig. 4 are well-known coastal wetlands in China. YC is located in Yancheng City, Jiangsu Province, and adjacent to the South China Sea. It is a typical muddy coastal wetland on the west coast of the Pacific Ocean with the largest area and the most complete ecosystem in the world [43]. Due to its unique geographical environment, the wetland has become the main growing area for wetland vegetation such as reed and Suaeda. The ZY1-02D hyperspectral dataset for YC was collected on September 6, 2020, and the sampling survey was a combination of visual interpretation of high-resolution images and field surveys. Table I shows the detailed information for 16 types of ground objects.

TABLE I
Training and Testing Sample Information of the YC

Class number	Ground objects	Number of samples	
		Training	Testing
1	bare land(bal)	24	211
2	suaeda	27	242
3	dry land(drl)	15	130
4	beach	33	298
5	river	21	193
6	fish pond(fip)	92	828
7	idle land(idl)	65	585
8	reed	24	220
9	sea	208	1868
10	paddy field(paf)	303	2723
11	salt pan(sap)	90	811
12	pond	7	61
13	shallow sea(shs)	156	1402
14	ditch	33	295
15	spartina alterniflora(spa)	31	275
16	building(bui)	45	406
	Total	1174	10548

YRD is situated on the coast of the Bohai Sea in the northeastern part of Shandong Province, China [44]. It is the most complete, broadest and comprehensive warm temperate zone of young wetland ecosystem in China [45]. Natural wetlands such as river, reed, and saline wetlands account for about 68.4%, while the other is artificial wetlands such as pond and reservoir [46]. The acquisition date of the ZY1-02D hyperspectral data in YRD was June 28, 2020. There are 23 ground objects and the details of the dataset are shown in Table II. Similar to YC sample collection, samples are selected through field investigation and visual interpretation of Google Earth.

C. Hyperspectral Data Preprocessing

The ZY1-02D hyperspectral data (L1-A product) is preprocessed by ENVI. First, the Global Digital Elevation Model Version 2 and rational polynomial coefficient are used to correct the ZY1-02D hyperspectral data. Second, radiometric calibration is conducted to convert the digital number into a radiance value by a linear function. Third, the atmospheric correction is performed by MODTRAN5 radiative transfer model. Fourth, bad bands and bands contaminated by clouds or water are removed. Finally, the images are masked by the vector data of our study areas. The data size of YC and YRD is 1398×942 pixels and 1147×1600 pixels, respectively.

TABLE II
TRAINING AND TESTING SAMPLE INFORMATION OF THE YRD

Class number	Ground objects	Number of samples	
		Training	Testing
1	salt pan filter pool(sfp)	25	222
2	spartina alterniflora(spa)	19	168
3	dry pond(drp)	14	126
4	salt pan evaporation pool(sep)	30	270
5	suaeda	22	196
6	tamarix	13	114
7	sea	469	4225
8	salt pan(sap)	31	275
9	river	58	526
10	tidal ditch(tid)	7	60
11	mud flat(muf)	2	14
12	idle land(idl)	46	413
13	locust	11	100
14	ecological restoration pool(erp)	31	279
15	building(bui)	40	358
16	fish pond(fip)	12	112
17	pond	13	115
18	paddy field(paf)	51	457
19	bare land(bal)	9	78
20	soybean(soy)	7	64
21	cotton	33	299
22	maize	10	93
23	reed	31	279
	Total	984	8843

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Setup

The proposed HSI-TNT is evaluated in two independent experiments. The first experiments explores the effects of different parameters on the model, including model depth, spatial patch size, and percentage of training samples. Second, the classification (mapping) performance and training time of the proposed HSI-TNT are compared with several common and typical classification approaches, including SVM with radial basis function and stochastic gradient descent (SGD), contextual deep CNN (CD-CNN) [47], 3D-CNN [48], and spectral-spatial fully convolutional networks conditional random field (SSFCN-CRF) [49]. Moreover, vision transformer (ViT) [50] is also included, which is compared with HSI-TNT as ablation experiments, due to ViT is composed of multilayer transformer models, while HSI-TNT is composed of multilayer nested dual transformers.

The experiments are implemented on the PyTorch platform installed on the Windows 10 computer. The learning rate is initialized at 1e-4 and decays by 0.5 times after the epoch size (50 epochs in total) reached 25, and loss function is the cross-entropy loss function. Each experiment is repeated 5 times independently. The overall classification accuracy (OCA), average classification accuracy (ACA), Kappa coefficient (KC), and training time are employed to quantify the classification accuracy.

B. Effects of Different Parameters

1) Effects of Model Depth L: Model depth is a key parameter that controls HSI-TNT complexity. High complexity with bigger L may lead to overfitting and waste of resources, while lower L may result into underfitting. Therefore, it is necessary to find a compromise L that can guarantee classification accuracy and avoid underfitting at the same time. With other parameters fixed, the parameters L are set from 2 to 10 for both datasets with an interval of 2. In Fig. 5, the training time increase sharply as L increase, while the optimal OCA up to 95.57% and 93.69% for YC and YRD when L is set to 6, respectively. The experimental results further prove that a moderate L is sufficient for optimal classification accuracy with limited resources.

2) Effects of Spatial Patch Size of the Sample: For both data, patch sizes are set from $5 \times 5 \times d$ to $15 \times 15 \times d$ (d is the number of bands) with an interval of 2. Due to the model's restrictions on the input data, the input size must be a composite number. However, if the input size is a prime number, it could become a composite number by adding 0. For example, if the input patch size is $5 \times 5 \times d$, 0 is added to both right and bottom sides, and then it becomes $6 \times 6 \times d$. The OCA and training time with different spatial patch sizes are shown in Fig. 6. The results show that when the patch size is 9, the highest OCA of YC and YRD are 95.57% and 93.69%, respectively. As for the training time, slowly increased first and then rapid growth are observed, especially when patch size greater than 9. And when the OCA is optimal, the training time for YC and YRD is 119.32 s and 102.53 s, respectively.

3) Effects of Percentage of Training Samples: Training samples are critical to the classification results. However, in HSI, *in situ* sampling is time consuming and labor intensive, and if an appropriate training samples percentage could be determined, researchers could save time and manpower without losing classification accuracy. Therefore, we explore the relationship of training sample percentage and OCA and training time and the results are shown in Fig. 7. We could see that, as the percentage of training samples increases, the classification performance gradually improves first, and at the same time the training time increases almost linearly. But OCA remains stable when more training samples are involved (e.g., 10%, 12%, and 14%), which largely indicates the proposed HSI-TNT could have good performance even with little training samples (10%).

C. Classification Accuracy and Mapping Results

In the experiments, L and patch size on both data are set to 6 and 9 for HSI-TNT, respectively. Of these, 10% of the samples are randomly selected for model training, and the remaining samples are used for testing. For the two datasets, the proposed HSI-TNT outperforms the other six methods with highest ACA (88.75%, 81.25%), KC (94.86%, 91.62%), and OCA (95.57%, 93.69%). Table III shows the classification results of YC, and OCA is improved by 8.28% and 4.43% compared to SVM and SGD, respectively. As for KC, HSI-TNT is 2.44% higher than ViT, which proves the importance of local features for classification. Moreover, HSI-TNT can well identify paddy field, sea, and shallow sea, which are also better than other methods.

HSI-TNT achieves the optimal accuracy for 12 of 23 classes, especially for salt pan evaporation pool and river, which are fully identified, as shown in Table IV. In addition, HSI-TNT increased by 3.13% for salt pan filter pond compared to 3D-CNN and increased by 6.36% compared to SSFCN-CRF for building,

TABLE III
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS FOR YC

Class	SVM	SGD	CD-CNN	3D-CNN	SSFCN-CRF	ViT	HSI-TNT
bui	97.53±0.39	97.47±0.76	88.10±6.46	**97.76±1.37**	93.89±2.80	83.90±4.39	95.31±0.85
suaeda	88.43±2.31	89.58±2.33	72.75±19.57	78.54±17.04	76.20±9.18	76.11±6.86	**90.27±0.60**
river	9.75±10.44	56.63±14.37	60.13±12.08	92.96±4.80	81.97±5.67	**98.23±1.82**	94.59±5.74
beach	0±0	0.27±0.53	**87.89±8.74**	42.29±20.33	76.71±18.38	85.05±8.56	86.60±5.65
bal	**98.25±0.77**	96.66±3.32	59.81±33.97	95.83±1.32	97.82±1.91	90.71±2.79	89.29±4.60
paf	98.19±0.39	98.34±0.11	98.52±0.47	98.17±1.66	99.49±0.35	99.36±0.34	**100±0**
idl	**98.25±1.14**	96.80±1.19	96.31±2.75	98.10±0.80	98.12±2.09	97.27±1.41	96.22±0.13
fip	57.07±3.39	78.07±1.21	83.29±2.64	78.15±14.43	**92.37±0.54**	81.19±4.82	87.49±0.97
sea	97.17±0.63	99.23±0.23	99.90±0.10	99.20±0.22	99.99±0.02	99.82±0.24	**100±0**
drl	68.31±18.81	69.90±6.23	58.24±22.54	**86.33±18.63**	85.19±5.91	67.23±3.02	78.62±5.85
sap	77.40±2.43	94.68±0.90	95.51±1.88	**96.14±0.72**	77.93±9.52	92.99±4.50	95.95±1.58
reed	86.95±0.34	**97.55±1.48**	72.56±25.38	82.32±13.31	84.00±17.59	84.93±7.62	86.12±1.37
shs	89.28±0.23	90.51±0.23	97.87±1.38	92.71±1.66	97.27±1.29	96.41±2.69	**100±0**
pond	0±0	0±0	0±0	37.50±16.50	4.26±6.19	28.85±6.36	**40.29±5.62**
spa	**95.75±0.91**	94.76±0.84	83.42±7.95	90.74±9.32	93.24±10.81	85.31±4.01	90.55±1.51
ditch	45.14±1.76	57.31±2.61	48.10±9.91	76.32±13.58	73.83±4.27	86.71±2.64	**88.44±2.63**
ACA	69.22±2.75	76.11±2.27	75.15±9.74	83.94±8.48	83.27±6.03	84.63±3.88	**88.73±2.32**
KC	85.15±0.70	89.67±0.58	90.46±0.64	91.80±2.81	92.81±1.39	92.42±0.58	**94.86±0.26**
OCA	87.29±0.59	91.14±0.50	91.79±0.55	92.95±2.40	93.81±1.19	93.46±0.50	**95.57±0.22**

TABLE IV
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS FOR YRD

Class	SVM	SGD	CD-CNN	3D-CNN	SSFCN-CRF	ViT	HSI-TNT
reed	23.49±4.70	36.79±6.80	40.25±8.48	54.36±11.84	**72.47±14.76**	56.49±5.61	67.89±5.91
spa	90.12±2.77	79.91±4.98	70.20±12.11	**94.09±4.25**	91.31±11.42	82.50±5.13	90.36±5.34
sfp	0±0	66.86±6.62	51.69±8.12	91.54±6.33	33.15±25.24	88.83±2.00	**94.67±2.14**
sep	62.59±7.32	79.90±6.72	39.53±22.50	93.81±2.20	95.26±6.91	97.63±1.81	**100±0**
drp	57.87±4.93	23.36±28.63	6.60±9.30	87.49±9.26	**93.33±1.47**	67.62±8.82	78.57±7.63
tamarix	47.20±24.76	62.88±2.58	22.04±27.01	58.59±29.66	**84.21±12.87**	55.44±11.06	72.81±2.78
sap	99.56±0.09	**99.71±0.15**	97.13±2.87	99.52±0.65	96.15±2.77	96.58±0.85	98.84±0.84
suaeda	83.46±2.29	78.32±1.86	76.02±9.27	94.24±3.05	96.23±3.50	94.59±2.00	**97.98±1.29**
river	98.21±0.33	98.71±0.15	98.14±0.79	98.69±0.37	99.81±0.24	99.54±0.31	**100±0**
sea	90.67±0.24	95.67±0.68	96.28±0.74	97.51±1.06	99.03±1.38	99.53±0.27	**99.94±0.06**
muf	0±0	0±0	0±0	30.36±37.38	27.69±28.62	13.34±4.08	**41.67±18.26**
tid	0±0	1.74±2.51	1.31±1.61	52.24±16.14	58.00±14.88	66.33±8.05	**77.67±4.42**
idl	88.06±1.58	87.63±2.47	80.17±2.24	93.38±2.31	**96.27±2.79**	86.78±3.59	91.82±3.54
erp	18.53±7.27	67.56±4.78	61.68±2.93	86.04±11.00	**90.68±4.05**	80.65±5.42	88.60±3.96
locust	0±0	0±0	0±0	55.23±28.24	0±0	40.20±12.30	**57.78±12.67**
fip	0±0	8.23±16.45	0±0	77.81±7.80	**96.07±7.86**	76.04±3.25	82.88±7.30
pond	0±0	0±0	0±0	36.16±33.20	12.00±24.00	54.78±5.36	**65.30±9.20**
bui	82.41±3.72	79.30±3.64	73.53±8.07	91.08±6.29	90.06±6.11	91.23±2.22	**96.42±1.44**
bal	0±0	0±0	0±0	68.35±34.85	73.33±25.04	62.82±10.03	**77.95±7.23**
paf	86.29±0.32	87.13±0.42	85.12±0.74	87.90±0.97	**99.30±1.40**	93.61±3.19	94.62±0.90
cotton	61.57±0.86	65.44±4.15	49.16±4.01	32.37±28.93	**91.04±8.42**	78.79±10.26	73.02±3.69
soy	0±0	0±0	0±0	32.50±20.20	32.06±29.58	24.76±10.07	**35.94±8.12**
maize	88.33±1.72	94.82±1.84	11.21±22.42	**97.95±2.76**	75.24±37.80	77.61±5.08	84.13±9.11
ACA	46.89±2.73	52.78±4.15	41.74±6.23	74.40±12.99	74.03±11.79	73.29±5.25	**81.25±4.86**
KC	73.13±0.62	79.43±0.79	73.58±2.19	86.98±2.83	88.49±0.52	87.75±0.54	**91.62±1.00**
OCA	80.97±0.43	84.93±0.54	80.54±1.57	90.27±2.05	91.35±0.41	90.79±0.41	**93.69±0.75**

TABLE V
TRAINING TIME (S) OF DEEP LEARNING METHODS FOR YC AND YRD

	CD-CNN	3D-CNN	SSFCN-CRF	ViT	HSI-TNT
YC	102.15	115.45	1311.82	56.94	119.32
YRD	85.00	95.42	960.02	47.49	102.53

which indicates the importance of global sequence features. The training time (see Table V) of HSI-TNT for two datasets is an acceptable result compared to other methods. Since the training samples of the YC are larger than that of YRD, the time of the YRD is shorter. Fig. 8 gives the OCA and training loss convergence curves of the two datasets, which depicts that good convergence can be achieved after 50 iterations with no overfitting.

In addition, to assess the statistically significant between HSI-TNT and the other six methods, we evaluate the different methods by McNemar's test [51]. The classification results of all samples of YC data are used for the test, and the $|Z|$-statistics results are obtained, as shown in Table VI. The null hypothesis is that HSI-TNT is not significantly differences with the other methods. The level of significance is 5% and the null hypothesis can be rejected if $|Z|>1.96$. Therefore, we can conclude that HSI-TNT is statistically significant with other methods from Table VI.

Figs. 9 and 10 further show the mapping results of different methods, which are in line with the results given in Tables III and IV. For example, other six methods easily misclassify reed as Suaede due to the similarity of them (elliptical in Fig. 9). Analogous results can be found in the rectangle in Fig. 10, where HSI-TNT can well recognize reed and Suaede, while other methods misclassify Suaede into fish pond and salt pan evaporation pool. In addition, combined with Tables III, IV and Figs. 9, 10, in the comparison of ablation experiments, HSI-TNT is superior

TABLE VI
STATISTICAL SIGNIFICANCE (|Z|) OF THE MCNEMAR'S TEST ON YC

	SVM	SGD	CD-CNN	3D-CNN	SSFCN-CRF	ViT
HSI-TNT	2.9129	7.8102	12.6557	2.4182	21.9721	8.0570

to ViT in both accuracy and mapping results, which indicates the critical contribution of inner T-Block and outer T-Block to HSI classification. In general, HSI-TNT has better applicability in large-scale HSI wetlands land cover classification and can better map the distribution of ground objects in coastal areas.

V. CONCLUSION

This article proposes a classification framework, named HSI-TNT, for mapping coastal wetlands using ZY1-02D HSI. The HSI-TNT utilizes position encodings and pixel-by-pixel unfold strategies to minimize the loss of spatial and spectral features; and then uses outer T-Block and inner T-Block to extract global and local features, also avoiding the loss of local information. Experimental results show that the method is robust and has good generalization ability, which can be applied to large-scale complex wetland scenes, and has the highest classification accuracy and moderate training time. We successfully applied this method to two independent coastal wetland areas, and in the future, we will investigate the generalization ability between different HSI scene data.

REFERENCES

[1] 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, 2020.

[2] H. Su, W. Yao, Z. Wu, P. Zheng, and Q. Du, "Kernel low-rank representation with elastic net for China coastal wetland land cover classification using GF-5 hyperspectral imagery," *ISPRS J. Photogramm. Remote Sens.,* vol. 171, pp. 238–252, 2021.

[3] M. J. McCarthy, K. R. Radabaugh, R. P. Moyer, and F. E. Muller-Karger, "Enabling efficient, large-scale high-spatial resolution wetland mapping using satellites," *Remote Sens. Environ.,* vol. 208, pp. 189–201, 2018.

[4] J. Wang *et al.*, "Estimation of aboveground vegetation nitrogen contents in the Yellow River estuary wetland using GaoFen-1 remote sensing data," *J. Coastal Res.,* vol. 102, no. SI, pp. 1–10, 2020.

[5] Y. Mao, D. L. Harris, Z. Xie, and S. Phinn, "Efficient measurement of large-scale decadal shoreline change with increased accuracy in tide-dominated coastal environments with Google Earth Engine," *ISPRS J. Photogramm. Remote Sens.,* vol. 181, pp. 385–399, 2021.

[6] N. J. Murray, R. S. Clemens, S. R. Phinn, H. P. Possingham, and R. A. Fuller, "Tracking the rapid loss of tidal wetlands in the Yellow Sea," *Front. Ecol. Environ.,* vol. 12, no. 5, pp. 267–272, 2014.

[7] P. Zuo, S. Zhao, C. A. Liu, C. Wang, and Y. Liang, "Distribution of spartina spp. along China's coast," *Ecological Eng.,* vol. 40, pp. 160–166, 2012.

[8] R. Liu *et al.*, "Ecosystem service valuation of bays in east China Sea and its response to sea reclamation activities," *J. Geographical Sci.,* vol. 30, no. 7, pp. 1095–1116, 2020.

[9] Z. Lin *et al.*, "OBH-RSI: Object-based hierarchical classification using remote sensing indices for coastal wetland," *J. Beijing Inst. Technol.,* vol. 30, no. 2, pp. 159–171, 2021.

[10] X. Wang *et al.*, "Tracking annual changes of coastal tidal flats in China during 1986–2016 through analyses of landsat images with Google Earth Engine," *Remote Sens. Environ.,* vol. 238, 2020, Art. no. 110987.

[11] A.-L. Balogun, S. T. Yekeen, B. Pradhan, and O. F. Althuwaynee, "Spatio-temporal analysis of oil spill impact and recovery pattern of coastal vegetation and wetland using multispectral satellite landsat 8-OLI imagery and machine learning models," *Remote Sens.,* vol. 12, no. 7, pp. 1225, 2020.

[12] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.,* vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[13] D. Stratoulias, H. Balzter, A. Zlinszky, and V. R. Tóth, "A comparison of airborne hyperspectral-based classifications of emergent wetland vegetation at Lake Balaton, Hungary," *Int. J. Remote Sens.,* vol. 39, no. 17, pp. 5689–5715, 2018.

[14] Y. Li *et al.*, "Progressive split-merge super resolution for hyperspectral imagery with group attention and gradient guidance," *ISPRS J. Photogramm. Remote Sens.,* vol. 182, pp. 14–36, 2021.

[15] P. Duan, Z. Xie, X. Kang, and S. Li, "Self-supervised learning-based oil spill detection of hyperspectral images," *Sci. China Technol. Sci.*, vol. 65, pp. 1–9, 2022.

[16] Z. Xie, J. Hu, X. Kang, P. Duan, and S. Li, "Multilayer global spectral-spatial attention network for wetland hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.,* vol. 60, Dec. 2022, Art. no. 5518913.

[17] W. Sun *et al.*, "A simple and effective spectral-spatial method for mapping large-scale coastal wetlands using China ZY1-02D satellite hyperspectral images," *Int. J. Appl. Earth Observ. Geoinf.,* vol. 104, 2021, Art. no. 102572.

[18] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.,* vol. 50, no. 4, pp. 1185–1198, Apr. 2012.

[19] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *IEEE Trans. Geosci. Remote Sens.,* vol. 42, no. 7, pp. 1552–1565, Jul. 2004.

[20] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.,* vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[21] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.,* vol. 43, no. 3, pp. 480–491, Mar. 2005.

[22] M. Partio, B. Cramariuc, M. Gabbouj, and A. Visa, "Rock texture retrieval using gray level co-occurrence matrix," presented at the 5th Nordic Signal Processing Symposium, Norway, vol. 75, Oct. 4-7, 2002.

[23] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.,* vol. 59, no. 9, pp. 7726–7738, Sep. 2021.

[24] Y. Qian, M. Ye, and J. Zhou, "Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features," *IEEE Trans. Geosci. Remote Sens.,* vol. 51, no. 4, pp. 2276–2291, Apr. 2013.

[25] L. Shen and S. Jia, "Three-dimensional Gabor wavelets for pixel-based hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.,* vol. 49, no. 12, pp. 5039–5046, Dec. 2011.

[26] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[27] A. O. B. Özdemir, B. E. Gedik, and C. Y. Y. Çetin, "Hyperspectral classification using stacked autoencoders with deep learning," in *Proc. 6th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2014, pp. 1–4.

[28] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors,* vol. 2015, 2015, Art. no. 258619.

[29] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by CNN denoiser," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 32, no. 3, pp. 1124–1135, Mar. 2021.

[30] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.,* vol. 56, no. 10, pp. 5893–5909, Oct. 2018.

[31] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT press, 2016.

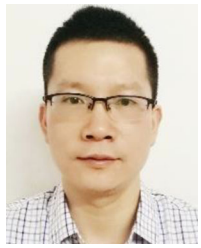
[32] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.,* vol. 12, no. 3, p. 582, 2020.

[33] 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.*, to be published, doi: 10.1109/TNNLS.2022.3149394.

[34] Y. Gao *et al.*, "Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5512615.
[35] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.,* vol. 60, Jul. 2022, Art. no. 5506812.
[36] M. Liang, L. Jiao, S. Yang, F. Liu, B. Hou, and H. Chen, "Deep multiscale spectral-spatial feature fusion for hyperspectral images classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 11, no. 8, pp. 2911–2924, Aug. 2018.
[37] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," 2021, *arXiv:2103.00112*.
[38] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.,* vol. 58, no. 1, pp. 165–178, Jan. 2020.
[39] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.,* vol. 13, no. 3, p. 498, 2021.
[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, p. 850, 2022.
[41] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, "Universal transformers," 2018, *arXiv:1807.03819*.
[42] W. Sun *et al.*, "A label similarity probability filter for hyperspectral image postclassification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 14, pp. 6897–6905, Jul. 2021.
[43] C. Wang *et al.*, "Diverse usage of waterbird habitats and spatial management in Yancheng coastal wetlands," *Ecol. Indicators,* vol. 117, 2020, Art. no. 106583.
[44] K. Ren, W. Sun, X. Meng, G. Yang, and Q. Du, "Fusing China GF-5 hyperspectral data with GF-1, GF-2 and Sentinel-2A multispectral data: Which methods should be used?," *Remote Sens.,* vol. 12, no. 5, p. 882, 2020.
[45] P. Cong, K. Chen, L. Qu, and J. Han, "Dynamic changes in the wetland landscape pattern of the Yellow River Delta from 1976 to 2016 based on satellite data," *Chin. Geographical Sci.,* vol. 29, no. 3, pp. 372–381, 2019.
[46] B. Cui, Q. Yang, Z. Yang, and K. Zhang, "Evaluating the ecological performance of wetland restoration in the Yellow River Delta, China," *Ecological Eng.,* vol. 35, no. 7, pp. 1090–1103, 2009.
[47] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.,* vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
[48] 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.
[49] Y. Xu, B. Du, and L. Zhang, "Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification," *IEEE Trans. Big Data,* vol. 6, no. 3, pp. 492–506, Sep. 2020.
[50] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[51] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogrammetric Eng. Remote Sens.,* vol. 70, no. 5, pp. 627–634, 2004.

Kai Liu received the B.S. degree in remote sensing science and technology from Shandong Normal University, Jinan, China, in 2020. He is currently working toward the master's degree in civil and hydraulic engineering with Ningbo University, Ningbo, China.

His current research interests include deep learning and hyperspectral image processing.



Weiwei Sun (Senior Member, IEEE) received the B.S. degree in surveying and mapping from Tongji University, Shanghai, China, in 2007, and the Ph.D. degree in cartography and geographic information engineering from Tongji University, Shanghai, China, in 2013.

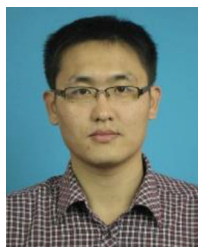
From 2011 to 2012, he was with the Department of Applied Mathematics, University of Maryland College Park, is a Visiting Scholar with the famous Prof. J. Benedetto to study on the dimensionality reduction of Hyperspectral Image. From 2014–2016, he was with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, working as a Postdoc to study intelligent processing in Hyperspectral imagery. From 2017 to 2018, he was with the Department of Electrical and Computer Engineering, Mississippi State University, also is a visiting scholar in hyperspectral image processing. He is currently a Full Professor with Ningbo University, Ningbo, China. He has authored and coauthored more than 80 journal papers. His current research interest includes hyperspectral image processing with machine learning.

Yijun Shao received the master's degree in engineering from the School of Materials Science and Chemical Engineering, Ningbo University, Ningbo, China, in 2011.

He is currently a Lecturer with the School of Materials Science and Chemical Engineering, Ningbo University, Ningbo, China. His research interests include engaged in ideological and political education research.

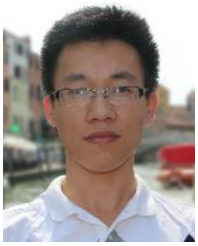
Weiwei Liu received the B.S. degree in agronomy from Shandong Agricultural University, Tai'an, China, in 2014, and the Ph.D. degree in agricultural remote sensing and information technology from Zhejiang University, Hangzhou, China, in 2020.

She is currently an Assistant Researcher with Ningbo University, Ningbo, China. Her current research interests include precipitation data fusion and remote sensing monitoring of agrometeorological disasters.



Gang Yang (Member, IEEE) received the M.S. degree in geographical information system from the Hunan University of Science and Technology, Xiangtan, China, in 2012, and the Ph.D. degree in cartography and geographical information engineering from the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, in 2016.

He is currently an Associate Professor with Ningbo University, Ningbo, China. His current research interests include missing information reconstruction of remote sensing image, cloud removal of remote sensing image, and remote sensing time-series products temporal reconstruction.



Xiangchao Meng (Member, IEEE) received the B.S. degree in geographic information system from the Shandong University of Science and Technology, Qingdao, China, in 2012, and the Ph.D. degree in cartography and geography information system from Wuhan University, Wuhan, China, in 2017.

He is currently an Associate Professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include remote sensing image fusion and quality evaluation.

Jiangtao Peng (Senior Member, IEEE) received the B.S. degree in information and computing sciences and M.S. degree in applied mathematics from Hubei University, Wuhan, China, in 2005 and 2008, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the Faculty of Mathematics and Statistics, Hubei University, Wuhan, China. His research interests include machine learning and hyperspectral image processing.

Dehua Mao received the B.S. degree in geographic information system from Northeast Forestry University, Harbin, China, in 2008, and the M.S. degree in cartography and geographic information system and Ph.D. degree in cartography and geographic information system from the University of Chinese Academy of Sciences, Beijing, China, in 2011 and 2014, respectively.

He is currently an Associate Researcher with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences. His research interests include remote sensing of wetland ecology, land cover change, and its environmental effects

Kai Ren received the B.S. degree in humanistic geography and urban-rural planning from Shanxi Normal University, Xi'an, China, in 2018, and the M.S. degree in humanistic geography from Ningbo University, Ningbo, China, in 2021.

He is currently a Research Assistant with the Faculty of Geographical Science and Tourism Culture, Ningbo University, China. His research interests include remote sensing data fusion and data analysis.