Spectrum-Induced Transformer-Based Feature Learning for Multiple Change Detection in Hyperspectral Images

Wuxia Zhang, Yuhang Zhang, Shiwen Gao, Xiaoqiang Lu, *Senior Member, IEEE*, Yi Tang, *Member, IEEE*, and Shihu Liu

***Abstract*— The multiple change detection (MCD) of hyperspectral images (HSIs) is the process of detecting change areas and providing "from–to" change information of HSIs obtained from the same area at different times. HSIs have hundreds of spectral bands and contain a large amount of spectral information. However, current deep-learning-based MCD methods do not pay special attention to the interspectral dependency and the effective spectral bands of various land covers, which limits the improvement of HSIs' change detection (CD) performance. To address the above problems, we propose a spectrum-induced transformer-based feature learning (STFL) method for HSIs. The STFL method includes a spectrum-induced transformer-based feature extraction module (STFEM) and an attention-based detection module (ADM). First, the 3D-2D convolutional neural networks (CNNs) are used to extract deep features, and the transformer encoder (TE) is used to calculate self-attention matrices along the spectral dimension in STFEM. Then, the extracted deep features and the learned self-attention matrices are dot-multiplied to generate more discriminative features that take the long-range dependency of the spectrum into account. Finally, ADM mines the effective spectral bands of the difference features learned from STFEM by the attention block (AB) to explore the discrepancy of difference features and uses the softmax function to identify multiple changes. The proposed STFL method is validated on two hyperspectral datasets, and their experiments illustrate the superiority of the proposed STFL method over the currently existing MCD methods.**

***Index Terms*— Attention, deep learning, hyperspectral images (HSIs), multiple change detection (MCD), transformer.**

I. Introduction

CHANGE detection (CD) is the process of recognizing whether the land cover of the same area has changed at different times. It is widely used in urban planning [1], [2], natural disaster detection [3], environmental detection, and other fields. Meanwhile, hyperspectral images (HSIs) contain hundreds of spectral bands and rich spectral information, and they are widely used in the fields of image classification [4], [5], [6], [7], [8], CD, and object recognition. Therefore, CD of HSIs has become a popular field in the computer vision field.

The CD is divided into binary CD (BCD) and multiple CD (MCD). The goal of BCD is only to detect whether there are changes in the land covers. The goal of MCD is not only to identify changes in the land covers but also to determine the specific types of changes, which provides valuable information to support fine urban planning and management. Therefore, this article will focus on the MCD methods for HSIs. The MCD methods for remote sensing images are usually divided into two categories: traditional machine-learning-based MCD methods and deep-learning-based MCD methods.

The traditional machine-learning-based MCD methods include change vector analysis (CVA) [9] and its extensions, active-learning-based and migration learning methods [10], differential feature-based methods [11], the hyperspectral unmixing method [12], [13], [14], and multi-objective-based genetic clustering methods [15]. However, most of the above methods use hand-crafted features, which cannot effectively express the deep semantics of HSIs or fully exploit their underlying information, leading to limited accuracy in detecting changes.

The deep-learning-based MCD methods can extract advanced features of HSIs in spatial and spectral aspects, determining change types more accurately. Moustafa et al. [16] trained MCD datasets using different kinds of UNet network structures and verified the performance of different kinds of networks using different loss functions. Saha et al. [17] processed diachronic images using prelearning convolutional neural networks (CNNs) and finally classified the samples by metric learning. Seydi and Hasanlou [18] generated training datasets using an image difference algorithm and a spectral unmixing approach, and finally used a CNN to generate multichange maps.

Although most of the deep learning methods mentioned above can extract deep features of HSIs well, they do not fully exploit the information in the spectral dimension. Due to the high spectral resolution and low spatial resolution characteristics of HSIs, spectral information plays a key role in the MCD of HSIs. Since a single spectral band cannot

Manuscript received 8 June 2023; revised 17 September 2023; accepted 28 September 2023. Date of publication 23 October 2023; date of current version 12 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62001378, in part by the National Science Fund for Distinguished Young Scholars under Grant 61925112, and in part by the Xingdian Talent Support Program for Young Talents under Grant XDYC-QNRC-2022-0518. *(Corresponding author: Wuxia Zhang.)*

Wuxia Zhang and Yuhang Zhang are with the Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: zhangwuxia@xupt.edu.cn).

Shiwen Gao is with the School of Information Engineering, Xi'an Eurasia University, Xi'an 710065, China.

Xiaoqiang Lu is with the College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China.

Yi Tang and Shihu Liu are with the School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650504, China.

Digital Object Identifier 10.1109/TGRS.2023.3325316

effectively represent the change region and there exists dependency and correlation between spectral bands, the dependency between spectral bands is critical to MCD tasks. However, most of the current deep-learning-based MCD methods do not fully explore the long-range dependency of HSIs in the spectral dimension, which can limit their performance in detecting changes. In addition, different land covers have varying degrees of sensitivity to different spectral bands during hyperspectral imaging. Therefore, certain spectral bands are more effective for detecting changes in certain land covers than others. However, current deep learning MCD methods do not adequately account for these differences, which can also impact their accuracy in detecting changes.

To address the above problems, we propose an MCD method for HSIs, named spectrum-induced transformer-based feature learning (STFL). The proposed STFL consists of a spectrum-induced transformer-based feature extraction module (STFEM) and an attention-based detection module (ADM). STFEM aims to extract discriminative features that consider the correlation of spectral bands. The purpose of the transformer encoder (TE) in STFEM is to calculate the self-attention matrices in the spectral dimension, which can explore the long-range dependency of the spectral bands. The discriminative features are acquired by performing the dot multiplication operation for deep features extracted from 3D-2D CNNs and the learned self-attention matrices. ADM aims to identify the multiple changes that take effective spectral bands for different land covers into account. The attention block (AB) in ADM tries to mine the effective spectral bands for different land covers, which helps improve the MCD performance.

The main contributions of our article can be summarized as follows.

1) We propose an STFEM, which captures the long-range dependency of the spectral bands by calculating self-attention matrices in the spectral dimension to extract discriminative features for multiple changes in HSIs.
2) We present an ADM, which mines the effective spectral bands of difference features extracted from STFEM by the AB to explore the discrepancy of difference features for multiple changes.
3) We design a compound loss function based on the cross-entropy loss function, which considers the difference information or its implicit information of the bottom, middle, and top layers during network training.

II. Related Work

A. Multiple CD

BCD just detects the areas in bitemporal images that have changed or remained unchanged, while MCD can not only locate changed areas but also identify changes in land-cover classes. Hence, MCD can acquire the "from–to" change information, providing the possibility to track the trajectory of changes [19], [20]. The MCD methods can be categorized into two distinct categories: traditional MCD methods and deep-learning-based MCD methods.

The traditional MCD methods are mainly based on CVA methods, which can be considered extensions of CVA. Bovolo proposed the C^2VA method, which is a classic CVA extension method [21]. It first introduced the vectorial angle to change the direction angle from $[0, 2\pi]$ to $[0, \pi]$ and then used the K-means clustering method to reduce the computation time. Hierarchical spectral CVA, proposed by Liu et al. [22], was an unsupervised CVA extension method that can analyze the hierarchical spectral changes at various levels. This approach provided the possibility to better simulate the complex land covers. Marinelli et al. [23] presented an unsupervised MCD method based on binary hyperspectral change vectors. It first encoded the initial spectral change vector as a binary code with the salient change information, and then a tree diagram was used to the learned binary spectral change vector code to analyze the change classes. Robust CVA, proposed by Thonfeld et al. [24], reduced spurious changes by considering pixel neighborhood effects.

The deep-learning-based MCD methods use deep neural networks to extract powerful and discriminative features to enhance the CD performance. Song et al. [25] presented an end-to-end MCD for HSIs, which combined a convolutional long short-term memory and a 3D fully convolutional network to preserve the spatial–spectral temporal information. Guo et al. [26] proposed a Siamese global learning framework for MCD with a backbone network composed of two UNet networks sharing weights. The features extracted from the backbone network were multiplied with a binary CD mask generated by an encoder. Finally, a multiclass change map was generated by the softmax function. Zhao et al. [27] used recurrent neural network (RNN) to mine the spectral information of the image and combined it with CNN to extract the spatial information of the image. The task of CD is finally accomplished by fusing the extracted features. Seydi and Hasanlou [18] first generated a binary change map by a series of algorithms. After that, a training dataset was generated by the image difference and spectral decomposition, which was subjected to feature extraction and finally used to generate multiple change maps by binary change maps. The aforementioned methods are primarily based on direct classification (DC) methods. Unlike the previous methods which are based on DC methods, the methods proposed by Zheng et al. [28] and Xia et al. [29] were based on post classification comparison (PCC) methods. They first obtained multiple change maps by classifying features of different images, then compared the classified results of different images to identify changes in land-cover classes, and finally generated a "from–to" change map that highlights the differences. Due to the separate classification processes for each time period in PCC methods, PCC CD methods suffer from error accumulation issues. Hence, our proposed approach is based on the DC method, which integrates images from different times and reduces the impact of error propagation.

B. Transformer-Based CD Method

Transformers were initially widely used in the field of natural language processing (NLP). However, with the introduction of various transformer-derived architectures, they have been widely adopted in various fields, including computer vision and remote sensing. Notably, vision transformers (ViTs) have

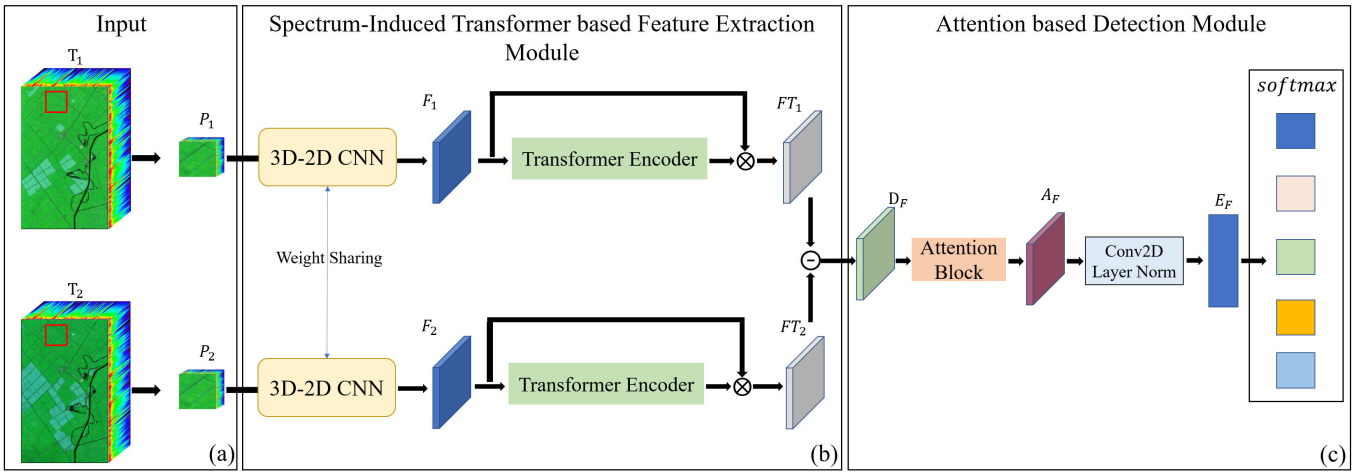

Fig. 1. Framework of the proposed STFL method. (a) Input port. (b) Spectrum-induced transformer-based feature extraction module consists of 3D-2D CNNs and a TE to extract the deep semantic features of HSIs. (c) ADM consists of an AB, a Conv2D operation, and a layer norm operation.

achieved success in the field of computer vision, demonstrating impressive results in tasks such as scene classification [30], [31], object detection [32], [33], image segmentation [34], [35], and CD [36], [37], [38] within the remote sensing domain. Unlike CNNs, ViTs [39] use self-attention mechanisms to learn long-range dependencies between features, allowing them to capture the global context more effectively.

The transformer has also been extensively used in CD for remote sensing images. Wang et al. [40] used a CNN to extract semantic features from the image, and then a transformer was used for temporal information interaction to obtain enhanced features. Subsequently, the extracted features were fed to the decoder to generate the change map. Chen et al. [36] used a CNN as the backbone network to extract features, which were subsequently subjected to contextual modeling in the spatial–temporal domain using the bitemporal image transformer (BIT). Zhang et al. [37] designed the SwinSUNet, which has a U-shaped structure and contains an encoder, a fusion module, and a decoder. The basic units are composed of swin transformer. Li et al. [38] proposed a hybrid model that combined the transformer and UNet to capture the long-term feature dependencies of images. All the above methods are focused on detecting binary changes.

C. Attention-Based CD Method

The attention mechanism enables the network to produce more representative features by weighing different parts of the input data differently and giving more importance to the areas that are more relevant to the specific task. Initially, attention mechanisms were used in machine translation applications and have been widely used in many different fields due to their ability to enhance the network performance. These fields include image classification [41], [42], image segmentation [43], [44], object detection [45], [46], and CD [47], [48], [49] of remote sensing images. Since remote sensing images primarily contain channel and spatial information, attention mechanisms can be divided into two categories: channel attention and spatial attention. They guide the feature extraction of the network from the channel dimension and the spatial dimension, respectively.

Chen and Shi [47] designed a spatial–temporal attention model for Siamese neural networks, which can calculate the position weights of two pixels at different times and generate more discriminative features. Peng et al. [48] extracted characteristics of land-cover changes by introducing upsampling spatial attention and upsampling channel attention to mine spatial contextual information. Jiang et al [49] used pyramid CNNs to extract deep features and incorporated a coattention mechanism to enhance the linkage between feature pairs, resulting in more discriminative features.

III. Method

The proposed hyperspectral MCD method based on STFL is shown in Fig. 1, which mainly contains two key modules: STFEM and ADM. The Siamese network is used as the backbone of the proposed STFL, which takes bitemporal HSI blocks at the same location as input, respectively. STFEM is capable of extracting discriminative features for multiple types of CD in HSIs, while ADM can explore the effective spectral bands of difference features learned from STFEM for multiple changes. The proposed method aims to explore the long-range dependency and discrepancy of the spectrum between image block pairs and finally achieves the detection of multiple changes in HSIs.

First, the HSIs at times T_1 and T_2 are block-cut one by one at the corresponding positions to obtain the image patch pairs $\boldsymbol{P_1}$ and $\boldsymbol{P_2} \in \mathbb{R}^{H\times W\times C}$ (H, W, and C represent the height, width, and channel of the image block, respectively). Second, $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$ are fed to 3D-2D CNNs' subnetwork to extract the deep features $\boldsymbol{F_1}$ and $\boldsymbol{F_2} \in \mathbb{R}^{m\times m\times d}$, respectively. Third, the learned deep features $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ are fed to the TE to achieve the corresponding weights $\boldsymbol{W_1}$ and $\boldsymbol{W_2} \in \mathbb{R}^{m\times m\times d}$. Fourth, $\boldsymbol{FT_1}$ and $\boldsymbol{FT_2} \in \mathbb{R}^{m\times m\times d}$ are obtained by performing dot multiplication operation on $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ with their corresponding weights $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$, respectively. Fifth, the difference feature $\boldsymbol{D_F} \in \mathbb{R}^{m\times m\times d}$ is obtained by subtracting $\boldsymbol{FT_1}$ and $\boldsymbol{FT_2}$. Finally, $\boldsymbol{D_F}$ is fed to the attention module to

TABLE I
PARAMETERS OF 3D-2D CNNS

Layer Name	Input Dim	Output Dim	KS	S
Conv3d	$1 \times C \times 7 \times 7$	$1 \times (C-6) \times 7 \times 7$	$7 \times 3 \times 3$	1
Conv3d	$8 \times (C-6) \times 7 \times 7$	$16 \times (C-10) \times 7 \times 7$	$5 \times 3 \times 3$	1
Conv3d	$16 \times (C-10) \times 7 \times 7$	$32 \times (C-12) \times 7 \times 7$	$3 \times 3 \times 3$	1
Conv2d	$(32 \times (C-12)) \times 7 \times 7$	$1024 \times 7 \times 7$	3×3	1
Conv2d	$1024 \times 7 \times 7$	$512 \times 5 \times 5$	3×3	1
Conv2d	$512 \times 5 \times 5$	$256 \times 3 \times 3$	3×3	1
Conv2d	$256 \times 3 \times 3$	$128 \times 3 \times 3$	3×3	1

learn more discriminative features which pay more attention to the changed areas. The softmax operation is then performed to generate the multiple change maps.

A. Spectrum-Induced Transformer-Based Feature Extraction Module

The STFEM consists of two main components, the 3D-2D CNNs' block and the TE.

1) 3D-2D CNNs: Roy et al. [50] proposed 3D-2D CNNs and good results were achieved in HSIs' classification. Since 3D convolution has an additional dimension compared with 2D convolution, it can extract deep features in the spectral dimension and exploit the relationship between spectral bands, providing a basis for the TE to explore spectral long-range dependency. However, using 3D CNN alone would result in a more complex network. By incorporating 2D CNN simultaneously, the network complexity can be greatly reduced. Therefore, we use a 3D-2D CNN hybrid approach for feature extraction.

As shown in Table I, the 3D CNN is used as the upsampling part to mine deep features in the spectral dimension of the image through three different layers of 3D convolutional kernels, while the 2D CNN is used as the downsampling part to fuse the features and reduce the network's complexity. Since the spectral information of HSIs is rich, it can be easily lost when using pooling operations. The CNN in the proposed STFL includes only convolution, not pooling operations. Unlike the article that proposed the 3D-2D CNNs, we only use it for the extraction of original features. In this way, the network structure is reasonably arranged to avoid the problem of numerous network parameters in the original paper. The bitemporal images' patches $\boldsymbol{P_1}$ and $\boldsymbol{P_2} \in \mathbb{R}^{H \times W \times C}$ are fed to the 3D-2D CNNs to obtain the original features $\boldsymbol{F_1}$ and $\boldsymbol{F_2} \in \mathbb{R}^{m \times m \times d}$, respectively.

2) Transformer Encoder: TE extracts the long-range dependency of spectrum by computing the weight matrix in the spectral dimension through a multi-head attention mechanism. Due to the large number of spectral bands in HSIs and the dependency and correlation between these bands, exploring the relationships and dependencies among spectral bands is beneficial for CD tasks. TE is based on the self-attention mechanism, which allows it to capture dependencies and relationships between different elements of an input sequence. TE treats the spectral bands as a sequence of data for HSIs. Therefore, we use TE to discover the correlation between spectral bands, which improves the effectiveness of band combination. As a result, the spectral bands of the extracted features more accurately reflect the spectral characteristics of the land covers.

As shown in Fig. 2(a), the TE is composed of L TE blocks (TEBs). Each TEB consists of layer norm, spectral-wise multi-head self-attention (S-MSA), and the feedforward operations as shown in Fig. 2(b). S-MSA, proposed by Cai et al. [51], [52], was the most important component and was first used in a spectral reconstruction task, achieving good results.

The structure of S-MSA is shown in Fig. 2(c). First, $\boldsymbol{X} \in \mathbb{R}^{mm \times d}$ are obtained by reshaping the input features $\boldsymbol{X_{in}} \in \mathbb{R}^{m \times m \times d}$. $\boldsymbol{W_V}$, $\boldsymbol{W_K}$, and $\boldsymbol{W_Q} \in \mathbb{R}^{mm \times Nd}$ (N is the number of self-attention heads) are obtained by the linear transformation in (1), where $\boldsymbol{W}$ represents the learnable weight, b denotes the bias parameter, and N indicates the number of heads in the multi-head self-attention (MSA) mechanism

$$\boldsymbol{W^{mm \times Nd}} = \boldsymbol{X^{mm \times d}} \times \boldsymbol{W^T} + b. \tag{1}$$

After that, it can be rearranged into N heads according to the spectral dimension to get value $\boldsymbol{V} = [V_1, V_2, \ldots, V_N]$, key $\boldsymbol{K} = [K_1, K_2, \ldots, K_N]$, and query $\boldsymbol{Q} = [Q_1, Q_2, \ldots, Q_N]$ matrices, where $V_i, K_i, Q_i \in \mathbb{R}^{mm \times d}, i \in [1, N]$. Spectral self-attention (SSA) is calculated as follows:

$$\boldsymbol{SSA_i} = \text{softmax}(\boldsymbol{K_i}\, \boldsymbol{Q_i^T}\, \delta_i) \tag{2}$$

$$\text{head}_i = \boldsymbol{SSA_i} \cdot \boldsymbol{V_i} \tag{3}$$

where $1 \leq i \leq N$ and $\boldsymbol{Q_i^T}$ represents the transpose of the matrix $\boldsymbol{Q_i}$. δ_i is a learnable parameter that is learned for N different spectral features to adapt $\boldsymbol{SSA_i}$. The ith head of MSA, denoted as head_i, can be calculated by (3), which differs from the traditional MSA in that it calculates a separate head_i for each $\boldsymbol{SSA_i}$.

Finally, $\boldsymbol{X_{out}} \in \mathbb{R}^{m \times m \times d}$ is calculated as follows:

$$\boldsymbol{X_{out}} = \text{Concat}(\text{head}_1, \ldots, \text{head}_n)\boldsymbol{W^T} + f(\boldsymbol{W_v}) \tag{4}$$

where $\boldsymbol{W}$ is the learnable weight, and $f(\cdot)$ is denoted as a 2D convolution operation. Since HSIs are sorted by wavelength along the spectral dimension, we use convolution to encode the position information of different spectral bands. As a result, the output feature $\boldsymbol{X_{out}}$ is obtained. Subsequently, $\boldsymbol{T W_L} \in \mathbb{R}^{m \times m \times d}$ can be acquired by passing L TEBs consecutively.

In the proposed STFEM, we aim to multiply the feature vectors obtained by TE as weights with the features extracted from 3D-2D CNNs (ensuring that the two feature dimensions are the same) and satisfy the MSA. However, the size of each head in S-MSA is $d_h = d/N$. Therefore, we need to modify $f(W_v)$ to perform feature fusion on W_v so that the fused features can be added up as biased and spliced features. By doing this, we can freely control the size of each head without controlling $d_h = d/N$. This approach makes the network structure more flexible and ensures that the feature vectors and weights have the same size.

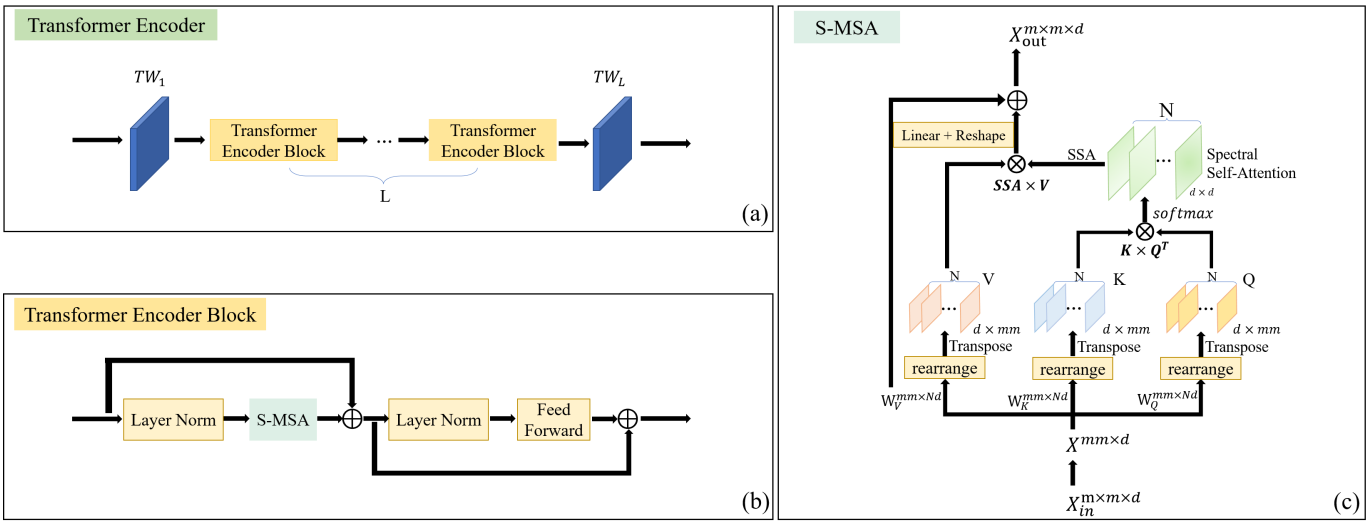

Fig. 2. Framework of the TE. (a) TE consists of a cascade of multiple TEBs. (b) TEB mainly consists of S-MSA and feed forward. (c) S-MSA captures the long-range dependency of spectrum by computing the weight matrix in the spectral dimension through a multi-head attention mechanism.

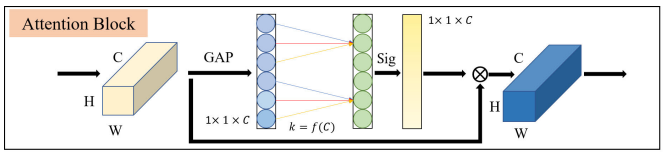

Fig. 3. AB enables cross-channel feature interaction through 1D convolution, enhancing the model's ability to capture dependency information between spectral features.

$\boldsymbol{FT_1}$ and $\boldsymbol{FT_2} \in \mathbb{R}^{m\times m\times d}$ are obtained from the following equation:

$$\boldsymbol{FT} = \boldsymbol{F} \times \boldsymbol{TW}. \tag{5}$$

We use the output of the TE as the weight instead of the output feature directly for two reasons: 1) Using the features of the TE as the weight, we can focus more on the spectral dependence information of important bands to better concentrate on the change region. 2) We aim to preserve the maximum amount of spectral information from the original features while avoiding the loss of spectral details in other bands by overemphasizing specific bands as a result of cascading multiple TEs.

B. Attention-Based Detection Module

The ADM mainly consists of the AB, a Conv2D operation, and a layer norm operation. The difference features $\boldsymbol{D_F} \in \mathbb{R}^{m\times m\times d}$ can be obtained by subtracting $\boldsymbol{FT_1}$ from $\boldsymbol{FT_2}$, which is fed to ADM to generate the multiple change maps. The AB applies the attention mechanism to difference features $\boldsymbol{D_F}$, which can enhance the CD results by selectively emphasizing the most relevant and effective spectral bands and suppressing noise and other irrelevant spectral bands.

The structure of AB [53] is shown in Fig. 3, which achieves cross-channel interaction of features by 1D convolution. This approach avoids the issues of dimensionality reduction and loss of feature information caused by the traditional channel attention mechanism. First, the input features are processed using the global average pooling operation (GAP). Subsequently, the 1D convolution operation and Sigmoid activation function (Sig) are performed to obtain the weights of each channel. Finally, $\boldsymbol{A_F}$ is obtained by multiplying the weights with the input features.

The feature $\boldsymbol{A_F} \in \mathbb{R}^{m\times m\times d}$ after passing the AB is used as the input feature. First, the deep semantic information in the feature vectors is extracted by 2D convolution with a kernel size of 3×3, and then downscaled to obtain 1D vectors. Finally, the final feature vector $\boldsymbol{E_F} \in \mathbb{R}^d$ is obtained by layer norm. CF(·) means the convolution and layer norm operation in the following equation:

$$\boldsymbol{E_F} = \mathrm{CF}(\boldsymbol{A_F}) \tag{6}$$

$$\text{Change Map} = \mathrm{softmax}(\boldsymbol{E_F}, \text{numclass}). \tag{7}$$

As shown in (7), the number of change types is set through the softmax operation to obtain multiple change maps.

C. Loss Function

The features in the middle layer of the deep network are also helpful for the CD task because the features extracted by the deep network are hierarchical and contain low-level, mid-level, and high-level semantic features. Therefore, the compound loss function is proposed in this article, which considers features extracted from 3D-2D CNNs as low-level semantic features, features that consider the TE information as mid-level semantic features, and the final high-level semantic features, simultaneously. The difference information and implicit information have been proven to be helpful for CD tasks. Therefore, the inputs to the loss function are all difference information or implicit information. The cross-entropy loss function, which provides better convergence and more uniform gradient values, is widely used in image classification. Hence, we adopt the cross-entropy loss function as the base loss function.

The presented compound loss function contains three terms, which are formulated as follows:

$$\text{Loss} = \delta_1 L_1 + \delta_2 L_2 + \delta_3 L_3 \tag{8}$$

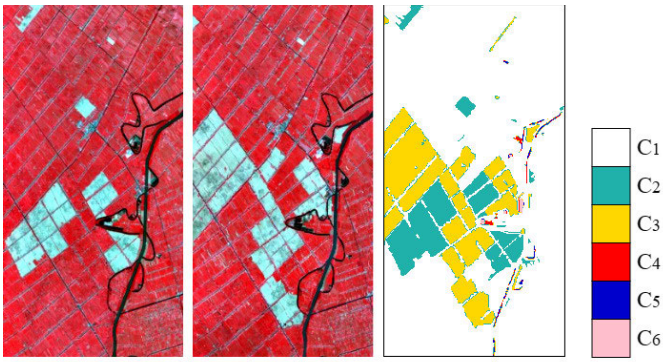

Fig. 4. Yangcheng dataset. (a) HSI was acquired in 2006. (b) HSI was acquired in 2007. (c) Ground truth.

where L_1–L_3 are the cross-entropy loss functions defined as follows:

$$L = -\frac{1}{\text{batch\_size}} \sum_{j=1}^{\text{batch\_size}} \sum_{i=1}^{n} y_{ij} \log \widehat{y_{ij}} \tag{9}$$

where $\widehat{y_{ij}}$ is the probability value of the network output, and y_{ij} represents the label information. $\widehat{y_{ij}}$ of L_1 is the $\boldsymbol{E_F}$. $\widehat{y_{ij}}$ of L_2 is the $\boldsymbol{D_F}$. The $\widehat{y_{ij}}$ of L_3 is the difference feature of $\boldsymbol{F_1}$ and $\boldsymbol{F_2} \in \mathbb{R}^{m \times m \times d}$.

In the presented compound loss function, we can simultaneously consider semantic features at different stages within the network, enabling better extraction of more discriminative spectral features.

IV. Experiment

In this section, we will conduct comparison experiments, parameter setup, and ablation studies on STFL using two hyperspectral datasets. First, we will introduce the hyperspectral dataset and the experiment setup separately. Second, we will investigate the performance of the proposed STFL through comparison experiments. Third, the optimal penalty parameter of the loss function will be determined through parameter setup. Finally, the effectiveness of the TE, the AB, the compound loss function, image block size and the 3D-2D CNNs will be explored via the ablation study.

A. Datasets

1) Yangcheng Dataset [19], [54]: As shown in Fig. 4, this dataset consists of bitemporal HSIs taken using the Hyperion sensor mounted on the EO-1 satellite in the wetland agricultural area of Ancheng, Jiangsu Province, China. The images were taken on (a) May 3, 2006 and (b) April 23, 2007, each with a size of 220 × 430 pixels. This dataset was obtained after preprocessing [55], with a total of 132 usable bands filtered out from the original 242 bands. The main types of changes include the below.

1) unchanged (C_1);
2) soil to vegetation (C_2);
3) vegetation to soil (C_3);
4) soil to water (C_4);
5) water to vegetation (C_5);
6) water to bare land (C_6).

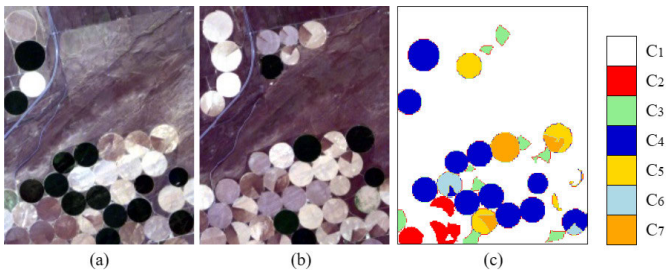

Fig. 5. Benton dataset. (a) HSI was acquired in 2004. (b) HSI was acquired in 2007. (c) Ground truth.

2) Benton Dataset [19], [54]: As shown in Fig. 5, this dataset consists of bitemporal HSIs taken using the Hyperion sensor mounted on the EO-1 satellite in the irrigated farmland area of Benton County, Oregon, USA. The images were taken on (a) May 1, 2004 and (b) May 8, 2007, each with a size of 180 × 225 pixels. The two images were obtained after preprocessing [55], with a total of 159 bands filtered from the original 242 bands. The land-cover changes (C_2–C_7) in this dataset mainly consist of class shifts between crops, bare soil, soil moisture changes, and changes between vegetation and water content. C_1 is indicated as unchanged pixels.

B. Experimental Setup

1) Training and Test Setup: The two datasets with the size of $H \times W \times C$ are processed pixel by pixel using a 7×7 sliding window, generating patches of size $7 \times 7 \times C$, with $H \times W$ patches obtained. The label of the center pixel in each patch is viewed as the label of the entire patch. This is because each pixel in an HSI contains rich spectral information, and the change information for each pixel is influenced by the surrounding pixels. The model can better detect the target information by extracting the information around the target pixels, thus improving the accuracy of CD.

During the dataset partitioning process, a sample imbalance problem exists in the multiple-class dataset. The rules for generating the training set are as follows:

$$\begin{cases} 50\%, & \text{class samples} < 6000 \\ 10\%, & \text{class samples} \geq 6000. \end{cases} \tag{10}$$

All the remaining samples are used as a testing set to evaluate the model's performance. We used a NVIDIA GeForce GTX 3080Ti GPU for training and testing. Adam is used to optimize the network.

2) Evaluation Criteria: We used four key evaluation metrics to analyze the performance of the model: overall accuracy (OA), precision, kappa, and F1-score. Each of them will be described below. OA, representing the model's overall performance, is the proportion of samples accurately classified by the classifier out of the total number of samples. It is defined as follows:

$$\text{OA} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{11}$$

where TP means true positive, TN represents true negative, FP indicates false positive, and FN means false negative.

Precision measures the ability of the classifier to correctly identify positive cases, while the kappa coefficient is a measure of classification accuracy. They are defined by the following equations:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \tag{12}$$

$$\text{kappa} = \frac{\text{OA} - P_e}{1 - P_e} \tag{13}$$

$$P_e = \frac{(\text{TP}+\text{FP})(\text{TP}+\text{FN}) + (\text{FN}+\text{TN})(\text{FP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})^2}. \tag{14}$$

The F1-score is the weighted average of precision and recall, which can balance the tradeoff between precision and recall, reflecting the model's overall performance. It is defined as follows:

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \tag{15}$$

$$\text{F1} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}. \tag{16}$$

C. Comparison Results and Analysis

1) Competitors:

1) KPCA-mnet [56]: KPCA-mnet extracts spectral features using the Siamese network architecture, which consists of kernel principal component analysis (KPCA) convolutional stacking. Then, unsupervised threshold segmentation or clustering methods are used to detect changes.
2) Spectral and spatial attention network (SSAN) [57]: SSAN takes the difference map of HSIs as input and consists of multiple attention mechanism modules. Then, a fully connected classifier is used to obtain the final results.
3) Deep Siamese convolutional network (DSCN): DSCN is the most commonly used deep learning method for CD. It takes bitemporal images as input, extracts features from the images using CNN, and generates the final change map through feature fusion.
4) Fully convolutional Siamese-concatenation (FC-Siam-conc) [58]: The model is based on the UNet architecture and uses Siamese networks in the encoding layers. The two encoding streams are directly connected in the decoding step.
5) Fully convolutional Siamese-difference (FC-Siam-diff) [58]: This model is based on the UNet architecture, and the structure of the model encoding layer is the Siamese network. The decoders in the model connect the absolute values of their differences.
6) BIT [36]: BIT uses ResNet as the backbone network for feature extraction. These extracted features are then subjected to contextual modeling in the spatial–temporal domain using BIT. This process ultimately generates the change map.
7) Binary change-guided hyperspectral multiclass CD network (BCG-Net) [59]: BCG-Net is trained jointly to generate binary CD maps through the united unmixing module and temporal correlation module. Subsequently, the end-element class corresponding to the maximum abundance value is used as the class of this image in the two temporal phases. Finally, the multiclass change map is obtained by comparing change classes.
8) Prior-aware transformer (PA-Former) [60]: PA-Former uses a prior-feature extractor to capture a prior and deep features from bitemporal images. Afterward, contextual markers with spatio-temporal information are obtained from the a priori features and integrated them into the deep features using a PA-Former module.

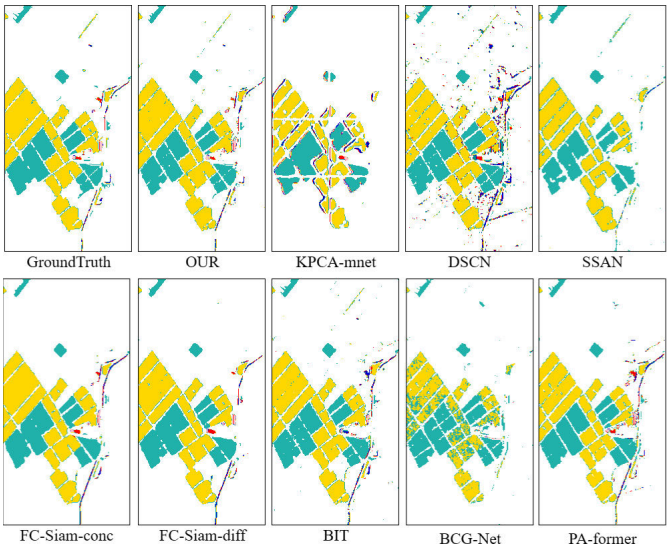

Fig. 6. Experimental result on the Yangcheng dataset.

2) Experiment on Yangcheng Dataset: The detection results of the Yangcheng dataset are shown in Fig. 6. It can be seen that the KPCA-mnet based on the traditional method has some miss detections and false detections. A miss detection occurs on the right side of the change map regarding the river changes (C_4–C_6) and a partial false detection occurs in C_2 and C_3 change types. Although the KPCA-mnet uses deep features, the original process of CD with KPCA does not involve feature extraction, so the extracted features may not necessarily meet the requirements of KPCA-mnet when detecting changes. Deep learning methods have a great improvement compared with traditional methods. However, there is a lot of noise in DSCN, and a serious misclassification occurs. There is a seriously miss detection in SSAN because C_4 and C_6 are not detected. The performance of FC-Siam-conc and FC-Siam-diff has lower miss detections and false detections, but the change map still shows a miss detection at the top right. Due to the small sample size of C_4–C_6, the above methods result in more severe omissions or misdetections for these classes. The above-mentioned deep-learning-based method does not extract the change features of the spectrum in depth, so there are more miss detections and false detections. While BIT outperforms the comparative methods, it produces some noise in the change map, leading to confusion between the categories C_4 and C_5. The PA-Former exhibits superior performance. However, it introduces some noise into the change map,

TABLE II
QUANTITATIVE RESULTS OF THE FOUR CRITERIA ON THE YANGCHENG DATASET

Method	OA	Precision	Kappa	F1
KPCA-mnet	0.8969	0.9143	0.7210	0.8949
SSAN	0.9613	0.9590	0.9030	0.9597
DSCN	0.9539	0.9746	0.8905	0.9629
FC-Siam-conc	0.9684	0.9702	0.9219	0.9690
FC-Siam-diff	0.9664	0.9672	0.6915	0.9666
BIT	0.9697	0.9715	0.9261	0.9703
BCG-Net	0.9248	0.9286	0.8116	0.9213
PA-former	0.9645	0.9679	0.9122	0.9655
STFL(ours)	**0.9759**	**0.9767**	**0.9407**	**0.9761**

causing confusion between categories C_4 and C_5. BCG-Net also exhibits confusion, particularly in categories C_2 and C_3, and it partially omits C_4–C_6. The proposed STFL method can extract the spectral features and make discriminations even for a small number of categories because of the more attention to the image spectral information. It can be seen from Fig. 6 that our proposed method shows better results.

Table II shows the results of the quantitative assessment of the compared methods and our method. The four metrics of KPCA-mnet based on traditional methods are lower than other deep-learning-based methods. SSAN, DSCN, and BCG-Net have poorer metrics among the deep-learning-based methods, while FC-Siam-conc, FC-Siam-diff, BIT, and PA-former are more effective than other comparison methods but still lower than the proposed STFL method. The above-mentioned methods do not focus on the long-range dependence of the spectrum of HSIs. The OA, precision, kappa, and F1 values of the proposed STFL method are 0.9759, 0.9767, 0.9407, and 0.9761, respectively, which are higher than the other comparison methods. Because the proposed STFL method pays more attention to the long-range dependencies of the spectrum of HSIs, the spectral information of the image is fully exploited.

3) Experiment on Benton Dataset: Fig. 7 depicts the detection results of the Benton dataset. The experimental results of KPCA-mnet based on the traditional method are poor. Regarding the deep-learning-based method, there are more false detections in SSAN, especially in the upper left of the change map. FC-Siam-conc and FC-Siam-diff have better results but still exhibit misclassification on C_2. DSCN provides the best results among the comparison methods but still has confusion between the two categories C_5 and C_7. The BIT method shows some miss detections in the lower left corner of the change map, and it confuses categories C_5 and C_7. The PA-Former method also exhibits miss detections in the lower left corner of the change map and confusion between categories C_5 and C_7. BCG-Net tends to confuse C_4 and C_6 categories and fails to detect C_2. The proposed STFL method enables the model to focus more on features of the changed region and suppress features of the unchanged

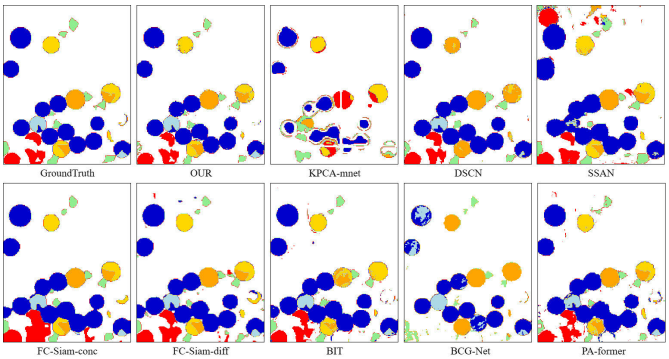

Fig. 7. Experimental result on the Benton dataset.

TABLE III
QUANTITATIVE RESULTS OF THE FOUR CRITERIA ON THE BENTON DATASET

Method	OA	Precision	Kappa	F1
KPCA-mnet	0.8304	0.8214	0.5179	0.8119
SSAN	0.8993	0.9408	0.7067	0.9088
DSCN	0.9620	0.9680	0.8635	0.9588
FC-Siam-conc	0.9451	0.9613	0.8229	0.9497
FC-Siam-diff	0.9453	0.9562	0.8218	0.9481
BIT	0.9505	0.9596	0.8332	0.9517
BCG-Net	0.9235	0.7764	0.8116	0.9213
PA-former	0.9576	0.9626	0.8548	0.9587
STFL (ours)	**0.9824**	**0.9824**	**0.9344**	**0.9818**

region through the spectral attention mechanism. Moreover, the spectral difference features are extracted from features of the changed region to explore the multi-change difference. Therefore, the results obtained by the STFL method are more accurate compared with other comparison methods.

Table III shows the quantitative results of the different methods. It can be seen that the detection results of KPCA-mnet based on the traditional methods, as well as SSAN and BCG-Net based on deep learning, are not satisfactory, because the lower precision and kappa coefficient indicate higher false and miss detection rates. DSCN, FC-Siam-conc, FC-Siam-diff, BIT, and PA-former have good results, but the metrics are still lower than our proposed STFL method. The STFL method achieves the best results in all the metrics with 0.9824, 0.9824, 0.9344, and 0.9818 for OA, precision, kappa, and F1, respectively. Because the AB in the ADM of the proposed STFL method can explore certain spectral bands that are more effective for detecting changes in certain land covers, which helps improve the CD performance of HSIs.

4) Runtime Cost Analysis: To ensure fairness in measuring training time and testing time, the dataset partitioning strategy is kept consistent for all the methods. Since KPCA-mnet is not accelerated with CUDA in the experimental platform, it is not involved in the comparison. BCG-Net provides only the total time overhead, hence lacks a separate testing time.

TABLE IV
EXPERIMENTAL RESULTS OF TRAINING AND TEST TIMES

Dataset	Time(s)	KPCA-mnet	SSAN	DSCN	FC-Siam-conc	FC-Siam-diff	BIT	BCG-Net	PA-former	STFL(ours)
YangCheng	Train	-	219.45	240	132.2	129.4	202.84	1452.75	123	312.4
Dataset	Test	-	19.85	25.97	15.67	15.23	12.02	-	17.2	27.18
Benton	Train	-	186.44	219.4	109.8	111.4	173.8	1300.38	161.2	297
Dataset	Test	-	7.85	11.79	7.22	6.83	5.04	-	6.99	12.18

As shown in Table IV, it can be seen that FC-Siam-conc, and FC-Siam-diff have the shortest training time and testing time. They rely on a basic CNN architecture with the fewest network parameters. Conversely, SSAN, BIT, and PA-former take slightly longer due to their integration of deep CNNs, various attention mechanisms, and transformers, which inherently increase the parameter count but also enhance the detection accuracy. DSCN, using the 3D-2D CNNs' framework, inherently has more network parameters, leading to longer training and testing times than FC-Siam-conc, FC-Siam-diff, SSAN, BIT, and PA-former methods. BCG-net takes the longest time. The proposed STFL method extracts deep features by 3D-2D CNNs and uses transform encoders and the attention mechanism to focus on the spectral dependency of effective bands. Therefore, the training and testing times of the proposed STFL method are only less than BCG-Net. However, our method demonstrates a significant improvement in accuracy metrics compared with other methods in both the datasets. Although our method is not the best in terms of efficiency, it achieves the best accuracy.

D. Parameter Setup

The penalty parameters δ_1, δ_2, and δ_3 in equation (6) are designed to balance the contributions of L_1–L_3 to the total loss. The penalty parameters δ_1–δ_3 are adjusted to explore the best combination of penalty parameters δ_1–δ_3. The loss information of L_1 in the proposed compound loss function is the most important because it contains the dependence information of the spectrum and the effective band information in the deep semantic features. Therefore, the weight information of L_1 is more significant than that of L_2 and L_3. Hence, δ_1 is set to 1. The experiments on two datasets are conducted to verify the impact of different penalty parameters δ_2 and δ_3 on the CD performance as shown in Figs. 8 and 9.

From Figs. 8 and 9, we can obtain that when $\delta_1 = 1.0$, $\delta_2 = 0.5$, and $\delta_3 = 0.5$, the four evaluation metrics are the highest on both the datasets. L_1 extracts the high-level semantic features that carry rich semantic information to capture the global context of HSIs. L_2 extracts the middle-level semantic features in the middle layer that consider the TE information and can represent parts of objects or simple object shapes. L_3 extracts low-level features of HSIs and only focuses on the local information. Thus, the model detection performance is best when $\delta_1 = 1.0$, $\delta_2 = 0.5$, and $\delta_3 = 0.5$.

E. Ablation Study

1) Effectiveness of TE: To explore the effectiveness of TE in the STFEM, we remove the TE from Fig. 2(a) and denote it as "wo Transformer Encoder." The effectiveness of TE is verified on two datasets, and their experimental results are shown in Table V.

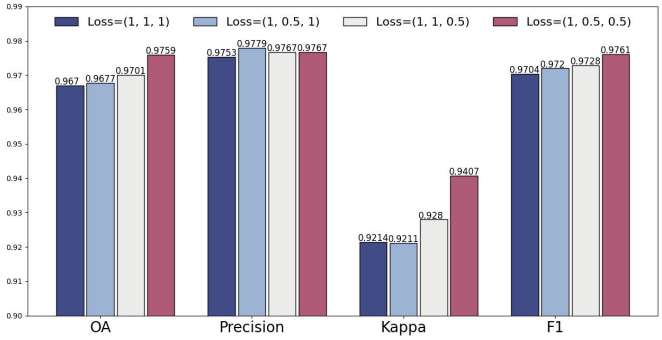

Fig. 8. Parameter setup on the Yangcheng dataset. Loss = (1, 1, 1) denotes $\delta_1 = 1$, $\delta_2 = 1$, and $\delta_3 = 1$, respectively.

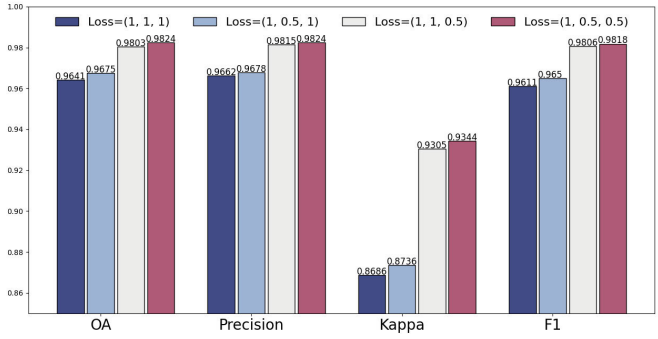

Fig. 9. Parameter setup on the Benton dataset. Loss = (1, 1, 1) denotes $\delta_1 = 1$, $\delta_2 = 1$, and $\delta_3 = 1$, respectively.

TABLE V
ABLATION STUDY ON TE AND AB

Dataset	Method	OA	Precision	Kappa	F1
Yang Cheng	wo Transformer Encoder	0.9639	0.9723	0.9140	0.9673
	wo Attention Block	0.9696	0.9769	0.9206	0.9725
	Both Have	0.9759	0.9767	0.9407	0.9761
Benton	wo Transformer Encoder	0.9725	0.9735	0.8987	0.9704
	wo Attention Block	0.9762	0.9772	0.9139	0.9756
	Both Have	0.9824	0.9824	0.9344	0.9818

Compared with the proposed "Both Have" model, the performance of the "wo Transformer Encoder" model is much lower than that of the "Both Have" model. This suggests that TE is important for the model to extract discriminative features for multiple classes of changes in HSIs.

2) Effectiveness of AB: To investigate the effectiveness of AB in the ADM, we remove the AB in Fig. 3 and denote it as

TABLE VI
ABLATION STUDY ON COMPOUND LOSS FUNCTION

Dataset	Loss($\delta_1, \delta_2, \delta_3$)	OA	Precision	Kappa	F1
Yang Cheng	(1,0,0)	0.9591	0.9650	0.9001	0.9616
	(1,1,0)	0.9671	0.9700	0.9172	0.9675
	(1,1,1)	0.9670	0.9753	0.9214	0.9704
Benton	(1,0,0)	0.9597	0.9656	0.8554	0.9592
	(1,1,0)	0.9623	0.9626	0.8513	0.9568
	(1,1,1)	0.9641	0.9662	0.8686	0.9611

"wo Attention Block." Ablation experiments are conducted on two datasets, and their detection results are shown in Table V.

Compared with the proposed model of "Both Have," the experimental metrics of "wo Attention Block" are lower than that of "Both have." This indicates that the AB significantly helps the model capture the effective spectral bands of different land covers, enabling the model to explore the spectral variability of multiple changes. Through the above experiments, we can conclude that when the model contains both TE and AB, it extracts the more discriminative features in the spectral dimension, which improves the CD performance.

3) Effectiveness of the Compound Loss Function: To verify the validity of the compound loss function, the penalty parameter δ_2 representing the contribution of the middle-level features and the penalty parameter δ_3 representing the contribution of the low-level features are set to 0 or 1, respectively. Then, ablation experiments are conducted on two datasets, and the CD performance is illustrated in Table VI.

As shown in Table VI, Loss $= (1, 0, 0)$ represents the calculation of loss in the compound loss function considering only high-level semantic features, but it exhibits the poorest CD performance on both the datasets. Loss $= (1, 0, 0)$ can capture global contextual information but might miss some important details and fine-grained information. Loss $= (1, 1, 0)$ takes into account both the high-level semantic information and mid-level features to calculate the loss of the compound loss function, leading to an increase in the four evaluation metrics. Because Loss $= (1, 1, 0)$ combines high-level semantic information with mid-level features, providing a balance between global context and local details when computing the loss. Loss $= (1, 1, 1)$ simultaneously considers low-level, mid-level, and high-level semantic features when calculating the loss of the compound loss function, and it achieves the best detection results. Because Loss $= (1, 1, 1)$ incorporates all the levels of features and can extract more discriminative features for CD of HSIs. However, the model performance is still lower than that of the best parameters shown in Figs. 8 and 9.

4) Effectiveness of the Image Block Size: The ablation experiment is conducted to verify the impact of varying sliding window sizes. The results are presented in Table VII, where we demonstrate the performance of three different sliding window sizes on two hyperspectral datasets. The size of the image block affects the CD performance. If the image block is too large, it might encompass multiple land-cover types and increase computational costs. If it is too small, there is a risk of missing crucial contextual information, making the block more noise-sensitive.

TABLE VII
EXPERIMENTAL RESULTS OF THREE DIFFERENT SLIDING WINDOW SIZES

Dataset	Image block size	OA	Precision	Kappa	F1
Yang Cheng	5×5	0.8694	0.7850	0.6171	0.8241
	7×7	0.9759	0.9767	0.9407	0.9761
	9×9	0.9628	0.9706	0.9110	0.9659
Benton	5×5	0.8719	0.8103	0.2609	0.8255
	7×7	0.9824	0.9824	0.9344	0.9818
	9×9	0.9761	0.9787	0.9153	0.9765

TABLE VIII
EXPERIMENTAL RESULTS OF 3D-2D CNNS

Dataset	Method	OA	Precision	Kappa	F1
Yang Cheng Dataset	with 3D-2D CNNs (STFL)	0.9759	0.9767	0.9407	0.9761
	wo 3D-2D CNNs	0.8643	0.9272	0.7211	0.8864
Benton Dataset	with 3D-2D CNNs (STFL)	0.9824	0.9824	0.9344	0.9818
	wo 3D-2D CNNs	0.9315	0.9192	0.7218	0.9134

As shown in Table VII, it can be seen that the worst results are acquired when the image block size is 5. The best results are achieved with a block size of 7. While a block size of 9 offers satisfactory performance, it is still inferior to the performance achieved with size 7. Consequently, we have chosen a block size of 7 for our experiments.

5) Effectiveness of the 3D-2D CNNs: To verify the effectiveness of the 3D-2D CNNs' feature extraction module, the ablation experiment is conducted. First, the entire 3D-2D CNNs' feature extraction module is removed, and the CD results on two datasets are shown in Table VIII. In Table VIII, "with 3D-2D CNNs" means that the structure of SFTL incorporates the 3D-2D CNNs' feature extraction module, while "wo 3D-2D CNNs" represents that the structure of SFTL does not include the 3D-2D CNNs' feature extraction module. It can be seen from Table VIII that the 3D-2D CNNs' feature extraction module positively influences the overall performance of our proposed STFL method. The removal of this module led to a decline in performance with respect to OA, kappa, precision, and F1 metrics. This indicates that the 3D-2D CNNs can mine rich semantic information, providing more useful information for the CD task.

V. CONCLUSION

In this article, we propose a spectral-induced MCD method for HSIs, named STFL method, which aims to capture the dependence between spectral bands and explore effective spectral bands of different land covers. STFL consists of two main modules: STFEM and ADM. STFEM extracts the deep features by 3D-2D CNNs to characterize the spatial and spectral information of HSIs. The spectral weights in the spectral dimension learned by TE are multiplied with the deep features extracted from 3D-2D CNNs to fully explore the long-range dependency of spectral bands and acquire discriminative features. The learned discriminative features of the bitemporal image block are subtracted and then fed to ADM to explore effective spectral bands for the changes in different land covers. The compound loss function is used to optimize the network. Finally, multiple change maps are

obtained by the convolution layer and the softmax function. The effectiveness of STFL is verified on two hyperspectral datasets for multiple classification. The experimental results show that the proposed STFL can achieve the best results in most cases. The proposed STFL method uses fixed-size sliding windows without accounting for varying shapes in changed regions. Although we alleviate data imbalance during dataset partitioning, it is not taken into account during training. In the future, we will focus on adaptive window design and incorporating mechanisms to handle sample imbalance during training.

REFERENCES

[1] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[2] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery," *Remote Sens.*, vol. 10, no. 7, p. 980, Jun. 2018.

[3] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, Nov. 2021, Art. no. 112636.

[4] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and LiDAR data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023.

[5] M. Zhang, X. Zhao, W. Li, Y. Zhang, R. Tao, and Q. Du, "Cross-scene joint classification of multisource data with multilevel domain adaption network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 6, 2023, doi: 10.1109/TNNLS.2023.3262599.

[6] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.

[7] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2817–2830, Jun. 2021.

[8] Y. Zhang, M. Zhang, W. Li, S. Wang, and R. Tao, "Language-aware domain generalization network for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501312.

[9] S. Liu, L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du, "Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4363–4378, Aug. 2015.

[10] X. Tong et al., "A novel approach for hyperspectral change detection based on uncertain area analysis and improved transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2056–2069, 2020.

[11] M. Zanetti and L. Bruzzone, "A theoretical framework for change detection based on a compound multiclass statistical model of the difference image," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1129–1143, Feb. 2018.

[12] S. T. Seydi, R. Shah-Hosseini, and M. Hasanlou, "New framework for hyperspectral change detection based on multi-level spectral unmixing," *Appl. Geomatics*, vol. 13, no. 4, pp. 763–780, Dec. 2021.

[13] H. Jafarzadeh and M. Hasanlou, "An unsupervised binary and multiple change detection approach for hyperspectral imagery based on spectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4888–4906, Dec. 2019.

[14] Q. Guo, J. Zhang, and Y. Zhang, "Multitemporal hyperspectral images change detection based on joint unmixing and information coguidance strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9633–9645, Nov. 2021.

[15] R. Pal, S. Mukhopadhyay, D. Chakraborty, and P. N. Suganthan, "Very high-resolution satellite image segmentation using variable-length multi-objective genetic clustering for multi-class change detection," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9964–9976, Nov. 2022.

[16] M. S. Moustafa, S. A. Mohamed, S. Ahmed, and A. H. Nasr, "Hyperspectral change detection based on modification of UNet neural networks," *J. Appl. Remote Sens.*, vol. 15, no. 2, Jun. 2021, Art. no. 028505.

[17] S. Saha, L. Kondmann, Q. Song, and X. X. Zhu, "Change detection in hyperdimensional images using untrained models," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11029–11041, 2021.

[18] S. T. Seydi and M. Hasanlou, "A new structure for binary and multiple hyperspectral change detection based on spectral unmixing and convolutional neural network," *Measurement*, vol. 186, Dec. 2021, Art. no. 110137.

[19] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[20] Q. Zhu, X. Guo, Z. Li, and D. Li, "A review of multi-class change detection for satellite remote sensing imagery," *Geo-Spatial Inf. Sci.*, pp. 1–15, Oct. 2022.

[21] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2196–2212, Jun. 2012.

[22] S. Liu, L. Bruzzone, F. Bovolo, and P. Du, "Hierarchical unsupervised change detection in multitemporal hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 244–260, Jan. 2015.

[23] D. Marinelli, F. Bovolo, and L. Bruzzone, "A novel change detection method for multitemporal hyperspectral images based on binary hyperspectral change vectors," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4913–4928, Jul. 2019.

[24] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 50, pp. 131–140, Aug. 2016.

[25] A. Song, J. Choi, Y. Han, and Y. Kim, "Change detection in hyperspectral images using recurrent 3D fully convolutional networks," *Remote Sens.*, vol. 10, no. 11, p. 1827, Nov. 2018.

[26] X. Guo, Q. Zhu, W. Deng, and Q. Guan, "A Siamese global learning framework for multi-class change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021v.

[27] H. Zhao, K. Feng, Y. Wu, and M. Gong, "An efficient feature extraction network for unsupervised hyperspectral change detection," *Remote Sens.*, vol. 14, no. 18, p. 4646, Sep. 2022.

[28] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 228–239, Jan. 2022.

[29] H. Xia, Y. Tian, L. Zhang, and S. Li, "A deep Siamese postclassification fusion network for semantic change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622716.

[30] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, p. 516, Feb. 2021.

[31] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.

[32] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608020.

[33] R. Xia et al., "CRTransSar: A visual transformer based on contextual joint representation learning for SAR ship detection," *Remote Sens.*, vol. 14, no. 6, p. 1488, Mar. 2022.

[34] H. Wang, X. Chen, T. Zhang, Z. Xu, and J. Li, "CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images," *Remote Sens.*, vol. 14, no. 9, p. 1956, Apr. 2022.

[35] L. Gao et al., "STransFuse: Fusing Swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.

[36] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[37] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[38] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[39] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23296–23308.

[40] G. Wang, B. Li, T. Zhang, and S. Zhang, "A network combining a transformer and a convolutional neural network for remote sensing image change detection," *Remote Sens.*, vol. 14, no. 9, p. 2228, May 2022.

[41] X. Tang et al., "Hyperspectral image classification based on 3-D octave convolution with spatial–spectral attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2430–2447, Mar. 2021.

[42] 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.

[43] X. Qi, K. Li, P. Liu, X. Zhou, and M. Sun, "Deep attention and multi-scale networks for accurate remote sensing image segmentation," *IEEE Access*, vol. 8, pp. 146627–146639, 2020.

[44] D. Xiao, Z. Wang, Y. Wu, X. Gao, and X. Sun, "Terrain segmentation in polarimetric SAR images using dual-attention fusion network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[45] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images," *Remote Sens.*, vol. 12, no. 3, p. 389, Jan. 2020.

[46] B. Cheng, Z. Li, B. Xu, X. Yao, Z. Ding, and T. Qin, "Structured object-level relational reasoning CNN-based target detection algorithm in a remote sensing image," *Remote Sens.*, vol. 13, no. 2, p. 281, Jan. 2021.

[47] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.

[48] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

[49] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, p. 484, Feb. 2020.

[50] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[51] Y. Cai et al., "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17481–17490.

[52] Y. Cai et al., "MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 744–754.

[53] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[54] S. Liu, Q. Du, X. Tong, A. Samat, and L. Bruzzone, "Unsupervised change detection in multispectral remote sensing images via spectral–spatial band expansion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3578–3587, Sep. 2019.

[55] S. Liu, Q. Du, X. Tong, A. Samat, H. Pan, and X. Ma, "Band selection-based dimensionality reduction for change detection in multi-temporal hyperspectral images," *Remote Sens.*, vol. 9, no. 10, p. 1008, Sep. 2017.

[56] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12084–12098, Nov. 2022.

[57] M. Gong et al., "A spectral and spatial attention network for change detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5521614.

[58] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[59] M. Hu, C. Wu, B. Du, and L. Zhang, "Binary change guided hyperspectral multiclass change detection," *IEEE Trans. Image Process.*, vol. 32, pp. 791–806, 2023.

[60] M. Liu, Q. Shi, Z. Chai, and J. Li, "PA-former: Learning prior-aware transformer for remote sensing building change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

Wuxia Zhang received the bachelor's degree in information display and opto-electronic technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the master's and Ph.D. degrees in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2012 and 2019, respectively.

From 2012 to 2016, she worked as a Software Engineer with Xi'an Huawei Technologies Company Ltd., Xi'an, China. Since 2019, she has been working as a Lecturer with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an. Her research interests include remote sensing and machine learning, especially remote sensing detection and deep networks with their applications in remote sensing.

Yuhang Zhang received the bachelor's degree in network engineering from the Xi'an University of Posts and Telecommunications (XUPT), Xi'an, China, in 2022, where he is currently pursuing the master's degree in computer science and technology.

His research interests include deep learning, and binary and multiclass change detection of remote sensing images.

Shiwen Gao received the bachelor's and master's degrees in telecommunications engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009 and 2012, respectively.

From 2012 to 2023, he worked as a Senior Software Engineer with Xi'an Huawei Technologies Company Ltd., Xi'an, China. His research interests include remote sensing detection and deep networks with their applications in remote sensing.

Xiaoqiang Lu (Senior Member, IEEE) is a Full Professor with the College of Physics and Information Engineering, Fuzhou University, Fuzhou, China. His research interests include pattern recognition, machine learning, hyperspectral image analysis, cellular automata, and medical imaging.

Yi Tang (Member, IEEE) is currently a Professor with the Department of Mathematics and Computer Science and the Key Laboratory of IOT Application Technology of Universities in Yunnan Province, Yunnan Minzu University, Kunming, China. His research interests include machine learning, statistical learning theory, image processing, and pattern recognition.

Shihu Liu is currently an Associate Professor with the School of Mathematics and Computer Sciences, Yunnan Minzu University, Kunming, China. His research interests include fuzzy sets, rough sets, graph data analysis, knowledge representation, and time series analysis.