CDFormer: A Hyperspectral Image Change Detection Method Based on Transformer Encoders

Jigang Ding, Xiaorun Li, and Liaoying Zhao

***Abstract*— Hyperspectral image (HSI) change detection (CD) has gained much attention in remote sensing. However, most deep-learning methods are restricted by a limited receptive field, without leveraging temporal information, and the need for many training samples. In this letter, we proposed a transformer encoder-based HSI CD framework called CDFormer. First, space and time encodings are added to the pixel sequence to guide transformers to exploit change information of space and time by the pixel embedding (PE) module. Second, the self-attention component of the transformer encoder module has a global space–time receptive field to mine the correlation and interaction between bi-temporal features, enhancing the utilization of temporal dependencies. Next, the multihead attention mechanism learns several attentions and extracts the joint weighted spatial–spectral–temporal features, which improves the feature discrimination ability of the changes. Finally, the detection result is predicted using a fully connected network. It is notable to mention that the proposed method only uses a few labeled samples to train the network. Experiments on two HSI datasets demonstrate that our proposed method can get effective performance in HSI CD.**

***Index Terms*— Change detection (CD), hyperspectral image (HSI), self-attention.**

I. Introduction

CHANGE detection (CD) is a significant technique that acquires different landscape information by analyzing the two images taken at the exact location at different times. Hyperspectral images (HSIs) make it possible to more accurately distinguish various objects at a fine spectral scale with its very high spectral resolution over a wide spectral wavelength range. HSI CD has been applied in land use, land cover change analysis, vegetation change analysis, and damage assessment.

Several approaches have been proposed to achieve better land cover CD in the literature. Generally, traditional CD algorithms can be categorized into image algebra-based and image transformation-based methods. Change vector analysis (CVA) calculates the magnitudes and directions of the bi-temporal images [1]. Later, some extensions of the CVA model have been proposed, for example, deep CVA (DCVA) [2]. Iterative reweighted multivariate alteration detection (IR-MAD) applies different weights to the linear combinations of the original variables at each iteration [3]. Hou et al. [4] proposed a novel three-order Tucker decomposition and reconstruction detector (TDRD) for HSI CD. The above methods usually have several limitations, such as inappropriate threshold selection, classification error, and model complexity.

In recent years, deep learning has exhibited good performance in remote sensing [5], [6], [7]. The convolutional neural network (CNN) is capable of extracting the representation features well. Song et al. [8] used a bidirectional reconstruction coding network and enhanced residual network to extract spectral and spatial features separately and fuse them. Saha et al. [9] employed an untrained CNN (UTCNN) model with some weight initialization strategy to extract semantic features. Zhao et al. [10] proposed a simplified 3-D convolutional autoencoder to extract the spatial–spectral simultaneously. These CNN-based methods are restricted by a limited receptive field and do not leverage the temporal dependency between bi-temporal images. However, spatial scope and temporal dependency are critical to identifying the change of interest in HSIs.

A recurrent neural network (RNN) is good at handling time-sequential data. Lyu et al. [11] utilized a long short-term memory (LSTM) network to learn the change rule from the information concerning changes, but spatial information has not been used. Moreover, some attempts have been made to learn joint spatial–spectral–temporal features from bi-temporal images by combining CNNs and RNNs. Song et al. [12] used a 3-D fully convolutional network and a convolutional LSTM (Re3FCN) to extract joint spectral–spatial–temporal features. Mou et al. [13] proposed a recurrent CNN (ReCNN), which extracted the spatial–spectral features from a 2-D CNN and fed the features into an RNN to learn the temporal dependency of changes. For these methods, the spatial–spectral information and temporal dependency are extracted individually, which cannot adequately capture the relative importance of each feature, and the complicated network structure makes choosing the optimum feature space challenging. Additionally, most deep-learning methods need many training samples, but collecting labeled training samples for CD tasks is time-consuming and difficult.

Manuscript received 30 June 2022; revised 3 August 2022; accepted 13 October 2022. Date of publication 25 October 2022; date of current version 14 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62171404 and in part by the Joint Fund Project of the Chinese Ministry of Education under Grant 8091B022120. *(Corresponding author: Xiaorun Li.)*

Jigang Ding and Xiaorun Li are with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 12010075@zju.edu.cn; lxyly@zju.edu.cn).

Liaoying Zhao is with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: zhaoly@hdu.edu.cn).

Digital Object Identifier 10.1109/LGRS.2022.3216878

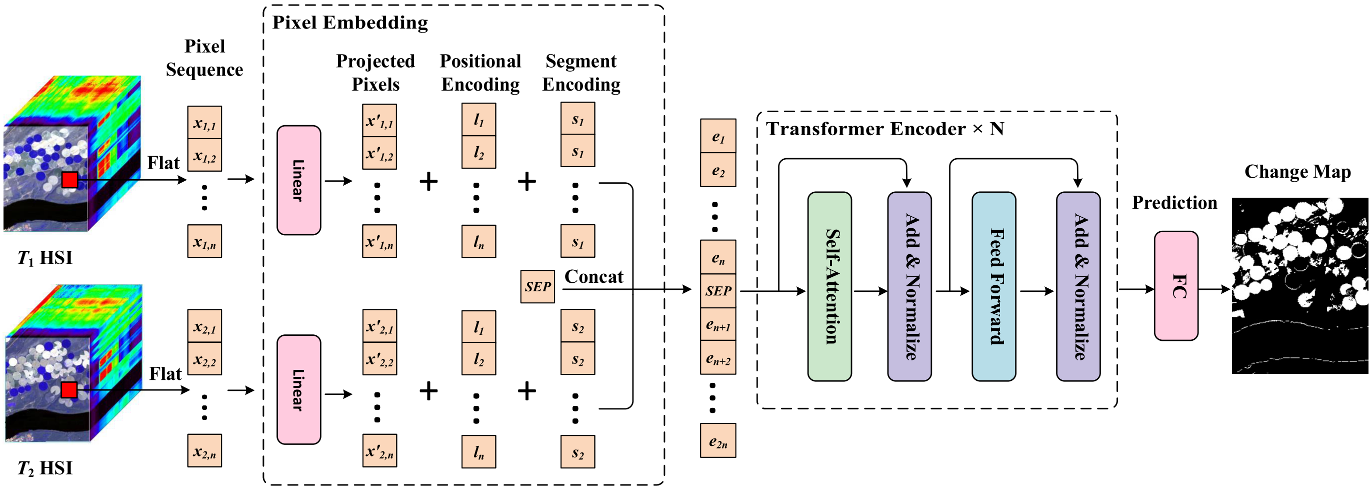

Fig. 1. Overview of our proposed method. First, two patches of HSIs are flattened into a pixel sequence, respectively. Then, the PE module transforms the sequences into an embedded sequence and adds spatial and temporal information to the sequence. Next, the sequence is fed into the transformer encoder module to exact joint weighted spatial–spectral–temporal features. Finally, a fully connected network is employed to generate the detection results.

To address these problems, we draw inspiration from neural language models. Neural language models can understand relationships between words and long-term dependencies across sentences. Actually, language models can be used in HSI CD with the proper conversion [14]. We analogize pixel sequences to sentences in natural language processing (NLP), because like word vectors, pixel vectors contain rich semantic and contextual information. Motivated by these analogies, we proposed a transformer encoder-based HSI CD framework called CDFormer. Transformers are designed to handle sequential input data and have made great success in NLP [15]. With a small number of labeled samples, CDFormer can effectively extract and integrate the joint weighted spatial–spectral–temporal features of the HSIs and enhance the leverage of temporal dependency and feature discrimination ability of the changes. The main contributions of our work can be summarized as follows.

1) The proposed method has a global space–time receptive field to capture spatial and temporal information simultaneously. The global receptive field mines the correlation and interaction between bi-temporal features, enhancing the leverage of temporal dependency.

2) The transformer encoder is introduced into HSI CD for the first time. The multihead self-attention (MHSA) of the transformer encoder captures the relative importance of features and extracts joint weighted spatial–spectral–temporal features, enhancing the feature discrimination ability of the changes.

3) The proposed method only needs a few labeled samples to train the network. It is beneficial to reduce the cost of sample labeling and improve the feasibility of automatic CD.

The remainder of this letter is organized as follows. In Section II, we elaborate on our proposed method. Experimental results on two datasets are presented in Section III. Finally, Section IV draws the conclusion.

II. Proposed Method

The architecture of CDFormer is shown in Fig. 1. First, two patches of the T_1 and T_2 HSIs are flattened into a pixel sequence, respectively. Then, the pixel embedding (PE) module transforms the pair of pixel sequences into an embedded pixel sequence with a predefined dimension and adds spatial and temporal information to the sequence. Next, the embedded pixel sequence is fed into the transformer encoder module, a multilayer transformer encoder. Through the MHSA mechanism, each pixel will attend to every pixel in an embedded pixel sequence. Finally, the learned features are fed into a fully connected network to obtain detection results.

A. Pixel Embedding

PE projects the pixel sequences into a new vector space, then adds spatial and temporal information to the sequences.

Let $X_t = (\boldsymbol{x}_{t,1}, \ldots, \boldsymbol{x}_{t,n}) \in \mathbb{R}^{n\times b}$ be the pixel sequence flattened from patches of one center pixel in $T_t (t = 1, 2)$ HSI, where n is the number of pixels and b is the number of hyperspectral bands. First, a fully connected network transforms each pixel sequence into a new feature space with a predefined dimension

$$X'_t = W^T X_t \tag{1}$$

where W is the learned weight of the linear layer. Through this operation, the transformer with the desired architecture can be used on HSIs with an arbitrary number of bands, and the training speed can be accelerated.

Second, we add positional encoding and temporal encoding to the sequences

$$E = X' + L + S \tag{2}$$

where $L = (\boldsymbol{l}_1, \ldots, \boldsymbol{l}_n)$ and $S = (\boldsymbol{s}_1, \boldsymbol{s}_2)$ denote the positional encoding and temporal encoding, respectively. Positional encoding is obtained by the Sinusoidal function [15]. The temporal encoding is a learned embedding to indicate whether a pixel belongs to T_1 or T_2. The encoding encodes the information about elements' relative or absolute position in space–time. Such position information can guide transformers to exploit change information of space and time.

Finally, the embedded pixel sequence is generated by concatenating the pair of pixel sequences and inserting

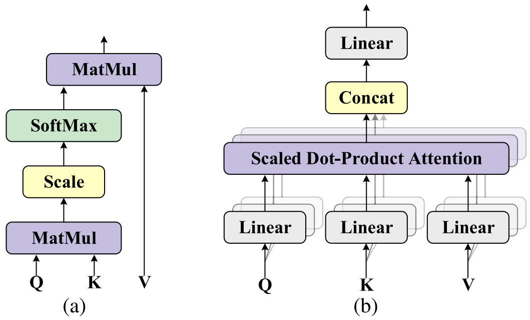

Fig. 2. (a) Scaled dot-product attention. (b) Multihead attention consists of several attention layers running in parallel.

the separator (SEP) token in the middle. The SEP token is a special separator token to separate T_1/T_2 pixel sequences.

B. Transformer Encoder

After obtaining the embedded pixel sequence for the input bi-temporal image, we then model the context of the sequence with the transformer encoder module, as shown in Fig. 1. Our motivation is that the global semantic relationships in space–time can be exploited by the transformer, thus producing the joint weighted spatial–spectral–temporal feature.

1) Multihead Self-Attention: The architecture of the MHSA mechanism is shown in Fig. 2. What the attention components do for each pixel in the output sequence is mapping the important and relevant pixels from the input sequence and assigning higher weights to these pixels, enhancing the feature discrimination ability of the changes. Our attention mechanism is scaled dot-product attention [15]

$$\text{Attention}(Q, K, V) = \mathbf{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V are matrices the queries, keys, and values packed into, and d_k is the dimension of the keys. The attention mechanism has a global space–time receptive field to integrate the different features oriented from each temporal feature, which leverages the temporal dependency.

MHSA performs multiple independent attention heads in parallel, and the outputs are concatenated and projected to obtain the final values, as shown in Fig. 2(b)

$$\text{MultiHead} = \mathbf{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{4}$$

where W^O represents the learned weight matrix, h is the number of attention heads, and

$$\text{head}_i = \text{Attention}(EW_i^Q, EW_i^K, EW_i^V) \tag{5}$$

with the projections using learned weight matrices $W^Q \in \mathbb{R}^{d\times d/h}$, $W^K \in \mathbb{R}^{d\times d/h}$, $W^V \in \mathbb{R}^{d\times d/h}$.

Multihead attention allows the model to jointly attend to change information from different representation subspaces at different positions, which can exploit the embedded pixel sequence's spatial, spectral, and temporal features at the same time and capture the relative importance of these features.

2) Feedforward: The feedforward module is applied to transform all heads' learned features further through fully connected layers. It consists of two linear transformations with rectified linear unit (ReLU) activation in between

$$\text{FFN}(\boldsymbol{x}) = \mathbf{max}(0, \boldsymbol{x}W_1 + \boldsymbol{b_1})W_2 + \boldsymbol{b_2}. \tag{6}$$

The parameter of the linear transformations across different positions is the same but is different from layer to layer.

3) Layer Normalization: Normalizing the activities of the neurons can reduce the internal covariate shift during the training process and make it faster. In this work, we use layer normalization.

C. Prediction

The prediction module adopts the average pooling to fuse the feature vector sequences to obtain the final joint weighted spatial–spectral–temporal feature. Then, a fully connected network is utilized to generate the final change map.

III. Experiments

A. Datasets

The performance of the proposed method is evaluated on two datasets. Both datasets are collected by the hyperion sensor aboard EO-1 and can be found in rslab.ut.ac.ir. The details are shown as follows.

1) Farmland: It covers an area of farmland in the city of Yancheng, Jiangsu, China, and is composed of 420 × 140 pixels. The two HSIs were taken on May 3, 2006, and April 23, 2007. After removing the noise and water absorption bands, there are 154 spectral bands used for CD.

2) Hermiston: This dataset belongs to an irrigated agricultural field in Hermiston, OR, USA, which was captured on May 1, 2004, and May 8, 2007, respectively. These two images consist of 307 × 241 pixels and 154 bands.

B. Experimental Settings

Our proposed method requires only a small number of labeled data to train the network. Therefore, 3% of the label samples in the above two datasets are randomly selected as training samples, and the samples of all methods are consistent. To demonstrate the effectiveness of the proposed method, six algorithms are selected as the compared methods: CVA, IR-MAD, TDRD, UTCNN, Re3FCN, and ReCNN. Furthermore, the metrics of overall accuracy (OA) and Kappa coefficient are exploited to evaluate the performance of the proposed CDFormer method.

Our models are implemented on Pytorch and trained 50 epochs using a single GTX 1080Ti GPU. We use the AdamW optimizer with a learning rate of 0.0001. The training batch size is set to 32. The dimension of the embedded sequence is 128.

C. Analysis of Model Depth and Attention Heads

The model depth is the number of transformer encoder layers in CDFormer. Both the model depth and the number of attention heads determine the complexity of CDFormer.

TABLE I

CD RESULTS OF CDFORMER IN DIFFERENT MODEL DEPTHS

Number Layers	1	2	3	4	5
Farmland	0.9505	0.9596	**0.9629**	0.9628	0.9608
Hermiston	0.9390	0.9472	0.9514	**0.9551**	0.9542

TABLE II

CD RESULTS OF CDFORMER IN DIFFERENT ATTENTION HEADS

Number Heads	1	2	4	8	16
Farmland	0.9618	0.9628	0.9604	**0.9647**	0.9634
Hermiston	0.9506	**0.9551**	0.9523	0.9512	0.9506

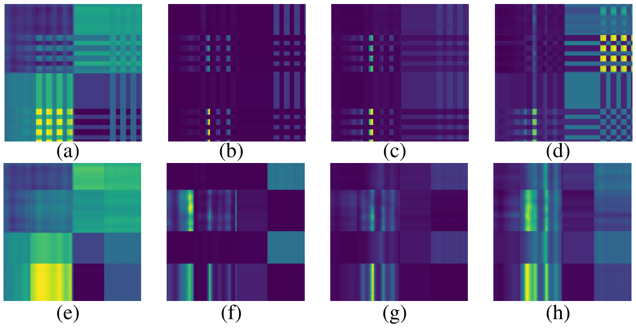

Fig. 3. Self-attention score visualization of CDFormer. (a)–(d) Attention score of head 1 from layers 1 to 4. (e)–(h) Attention score of head 2 from layers 1 to 4. Each attention score matrix is in the shape of 99×99 (including 98 pixels and 1 SEP token). The ith row of each attention matrix corresponds to the attention score over the whole pixel sequence of the ith pixel.

To find the smallest model depth and the number of attention heads without underfitting, we fix all other parameters and vary the model depths and the number of attention heads, respectively. The OA in different model depths is listed in Table I. The best model depths for the farmland and Hermiston datasets are 3 and 4. Table II shows the OA for various numbers of attention heads. The best number of attention heads for the two datasets is 8 and 2.

In this experiment, the pair of 7×7 square patches are used to train the model. The visualization of the attention score of different layers and heads is shown in Fig. 3. The attention score distribution of pixels for different heads in the same layer is obviously varied, implying that CDFormer jointly attends to information from various representation subspaces at each position. As the number of layers increases, the attention score becomes more diverse, suggesting that CDFormer learns more complex attention as the model depth grows.

D. Analysis of the Number of Training Samples

To further illustrate the superiority of CDFormer, we also conducted experiments with a different number of labeled training samples. We sampled 1%, 3%, 5%, 10%, and 20% labeled samples in strata to train Re3FCN, ReCNN, and CDFormer on the farmland dataset. The OA results are tabulated in Table III. CDFormer surpasses other methods in different numbers of training data.

TABLE III

ACCURACY COMPARISON AMONG DIFFERENT NUMBERS OF TRAINING SAMPLES ON FARMLAND DATASET

Number (%)	1	3	5	10	20
Re3FCN	0.9449	0.9556	0.9616	0.9672	0.9700
ReCNN	0.9231	0.9395	0.9455	0.9534	0.9604
CDFormer	**0.9455**	**0.9647**	**0.9662**	**0.9685**	**0.9770**

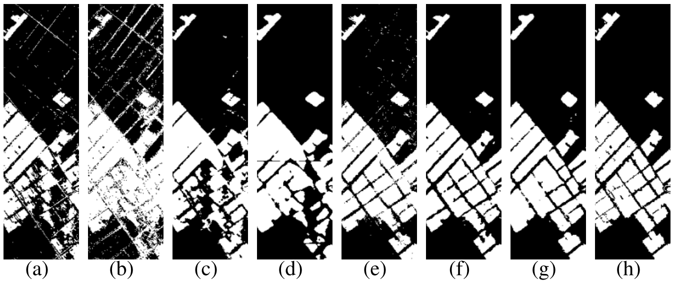

Fig. 4. Detection results for farmland. (a) CVA. (b) IR-MAD. (c) TDRD. (d) UTCNN. (e) ReCNN. (f) Re3FCN. (g) Proposed. (h) Ground truth.

TABLE IV

QUANTITATIVE COMPARISON AMONG DIFFERENT METHODS ON THE FARMLAND DATASET

	Unsupervised				Supervised		
	CVA	IR-MAD	TDRD	UTCNN	Re3FCN	ReCNN	Proposed
OA	0.8749	0.8287	0.8847	0.8860	0.9556	0.9395	**0.9647**
Kappa	0.6998	0.6368	0.7309	0.7316	0.8966	0.8606	**0.9185**

The good results can be owed to the following aspects: 1) the global space–time receptive field mines the correlation and interaction between the bi-temporal features, enhancing the utilization of temporal dependency and 2) the MHSA mechanism captures the relative importance of features and extracts the joint weighted spatial–spectral–temporal features, which enhances the feature discrimination ability of the changes. The above allows the model to mine more information about changes with only a few samples.

E. Results

1) Farmland: The CD results on the farmland dataset are illustrated in Fig. 4(a)–(g), and the evaluation indices are shown in Table IV. Among the three traditional methods, TDRD achieves the best results. The reason is that it considers spatial information and spectral information. This also confirms the importance of joint features. Although UTCNN is a deep-learning approach, it is slightly more effective than TDRD because it relies only on the weight initialization strategy. Compared to the above methods, there are significant improvements in all evaluation indicators achieved by supervised deep-learning methods. It is also noticed that these unsupervised methods have a large number of false-detected pixels in the lower middle area, which are correctly detected by supervised methods. And our proposed CDFormer has achieved the best performance compared to Re3FCN and ReCNN. For instance, the accurate increments on OA are 0.91% and 2.52%. It is possible that our proposed

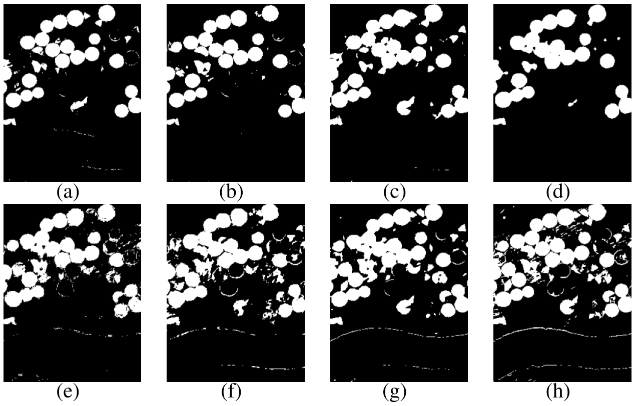

Fig. 5. Detection results for Hermiston. (a) CVA. (b) IR-MAD. (c) TDRD. (d) UTCNN. (e) ReCNN. (f) Re3FCN. (g) Proposed. (h) Ground truth.

TABLE V

QUANTITATIVE COMPARISON AMONG DIFFERENT METHODS ON THE HERMISTON DATASET

	Unsupervised				Supervised		
	CVA	IR-MAD	TDRD	UTCNN	Re3FCN	ReCNN	Proposed
OA	0.9200	0.9174	0.9285	0.9063	0.9308	0.9284	**0.9551**
Kappa	0.7410	0.7318	0.7778	0.6964	0.7993	0.7821	**0.8685**

method extracts joint weighted spatial–spectral–temporal features that can express the change rule better.

2) Hermiston: The results of the Hermiston dataset can be found in Fig. 5(a)–(g) and Table V. From the ground truth, the changes in this dataset are more complicated, and the change area contains a lot of irregular and small areas. The three traditional methods barely detect these changes, which is the main reason that affects their accuracy. For example, the changes in the river's edge are not detected well. The deep-learning method has the ability to learn more representative features and make up for the lack of accuracy of traditional methods. However, the overall performance of ReCNN and UTCNN is weaker than TDRD. The guess is that the 2-D CNN used by them is restricted by the limited receptive field. And our proposed CDFormer outperforms the best comparison method with improvements of 2.43% on OA and 0.0692 on Kappa. The reason may be that our proposed method's global space–time field can obtain richer contextual information to extract small irregular areas' features.

IV. CONCLUSION

In this letter, we proposed a transformer encoder-based HSI CD framework. The proposed CDFormer is primarily built on the MHSA mechanism that extracts and integrates the joint weighted spatial–spectral–temporal features of the HSIs, enhancing the leverage of temporal dependency and feature discrimination ability of the changes. Compared with the classical and state-of-the-art methods, the detection results of our proposed architecture outperformed them in binary CD with only a small number of labeled data.

REFERENCES

[1] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2007.

[2] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[3] A. A. Nielsen, "The regularized iteratively reweighted mad method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[4] Z. Hou, W. Li, R. Tao, and Q. Du, "Three-order tucker decomposition and reconstruction detector for unsupervised hyperspectral change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6194–6205, 2021.

[5] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[6] Y. Gao et al., "Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[7] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, "Graph information aggregation cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 60, pp. 1–14, 2022.

[8] B. Song, Y. Tang, T. Zhan, and Z. Wu, "BRCN-ERN: A bidirectional reconstruction coding network and enhanced residual network for hyperspectral change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[9] 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.

[10] C. Zhao, H. Cheng, and S. Feng, "A spectral–spatial change detection method based on simplified 3-D convolutional autoencoder for multitemporal hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[11] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, p. 506, 2016.

[12] 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.

[13] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[14] 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, Sep. 2020.

[15] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.