CSANet: Cross-Temporal Interaction Symmetric Attention Network for Hyperspectral Image Change Detection

Ruoxi Song, Weihan Ni, Wei Cheng, and Xianghai Wang

***Abstract*— Deep learning methods have been extensively applied to hyperspectral (HS) image change detection (CD) tasks and achieved promising performance. However, the beneficial joint spatial–spectral–temporal information provided by the HS images has not been fully used. Since the bi-temporal HS images are highly symmetric, in this letter, we propose a novel cross-temporal interaction symmetric attention network (CSANet), which can effectively extract and integrate the joint spatial–spectral–temporal features of the HS images, at the same time to enhance feature discrimination ability of the changes. Specifically, we propose a novel cross-temporal interaction symmetric attention (CSA) module to interact with the bi-temporal HS information, where self-attention is combined to enhance the feature representation ability of each temporal image, and the cross-temporal attention is utilized to integrate the difference features oriented from each temporal feature embedding. On this basis, we design a Siamese network structure equipped with the CSA to hierarchically extract the change information in a symmetric pattern. Experimental results on three public HS image CD datasets show that the proposed CSANet CD framework achieves a significant improvement when compared with the state-of-the-art (SOTA) methods. The source code of the proposed framework will be released at https://github.com/srxlnnu.**

***Index Terms*— Change detection (CD), cross-temporal interaction, hyperspectral (HS) image, self-attention.**

I. Introduction

With the rapid development of modern remote sensing techniques, remote sensing image change detection (CD) has become one of the most important means of monitoring the transition process of land-cover objects [1]. Hyperspectral (HS) images can not only depict the spatial size and distribution of the land-cover objects, but can also draw the corresponding approximately continuous electromagnetic spectral reflection curve for them. Therefore, the fine spectral resolution of HS images brings the possibility of capturing the subtle changes associated with the land-cover dynamic evolution process.

HS image CD can be categorized into early-stage approaches and deep learning-based approaches. In the early-stage HS image CD approaches, unsupervised feature extraction methods are applied to obtain the similarities between each HS image pixel. The automatic or manually selected threshold is then determined to obtain the changes. Representative methods include change vector analysis (CVA), principal component differential analysis (PCDA), three-order tucker decomposition and reconstruction detector (TDRD) [2]. However, these early-stage approaches are often designed to extract shallow-level features of the HS images, thus the CD accuracy is extremely limited. Meanwhile, hard threshold selecting strategies make the models lack the robustness of complex scenes.

In recent years, HS image CD has witnessed substantial breakthroughs in terms of both model accuracy and model efficiency [3]. Wang *et al.* [4] proposed a general end-to-end 2-D CNN HS image CD framework (GETNET), where the abundance of information of the HS images is combined with 2-D CNN to improve the CD accuracy. Zhan *et al.* [5] proposed a novel HS image CD framework based on spectral–spatial CNN with Siamese architecture, where the HS images are first fed into the Siamese CNN to extract the shallow level spectral–spatial vectors, then Euclidean distances of the two vectors are calculated to represent the similarity of the tensor pairs. To simultaneously extract the spatial–spectral features of the HS images, Zhao *et al.* [6] proposed a novel HS image CD framework based on a simplified 3-D convolutional autoencoder (S3DCAECD), and experimental results demonstrate that S3DCAECD can effectively reduce spectral redundancy of the HS images.

Existing deep learning-based HS image CD approaches have already achieved promising performance, however, they still facing the following limitations. First, existing approaches lack the utilization of joint spatial–spectral–temporal features, since the bi-temporal HS images are highly symmetric, and the inherent correlation within the image pair can be fully used by exploiting its symmetry cues. Meanwhile, in the modeling process, most of the existing methods often ignore the correlation and interaction between the multitemporal features, which can greatly improves the CD accuracy of the model [7].

In this letter, we design and construct the cross-temporal interaction symmetric attention network (CSANet) for HS image CD, which can effectively extract and integrate the joint spatial–spectral–temporal features of the HS images, and at the same time enhancing feature discrimination ability of

Manuscript received April 27, 2022; accepted May 25, 2022. Date of publication May 30, 2022; date of current version June 9, 2022. This work was supported by the National Natural Science Foundation of China under Grant 41971388. *(Corresponding author: Xianghai Wang.)*

Ruoxi Song and Xianghai Wang are with the School of Geography, Liaoning Normal University, Dalian 116029, China (e-mail: xhwang@lnnu.edu.cn).

Weihan Ni and Wei Cheng are with the School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China.

Digital Object Identifier 10.1109/LGRS.2022.3179134

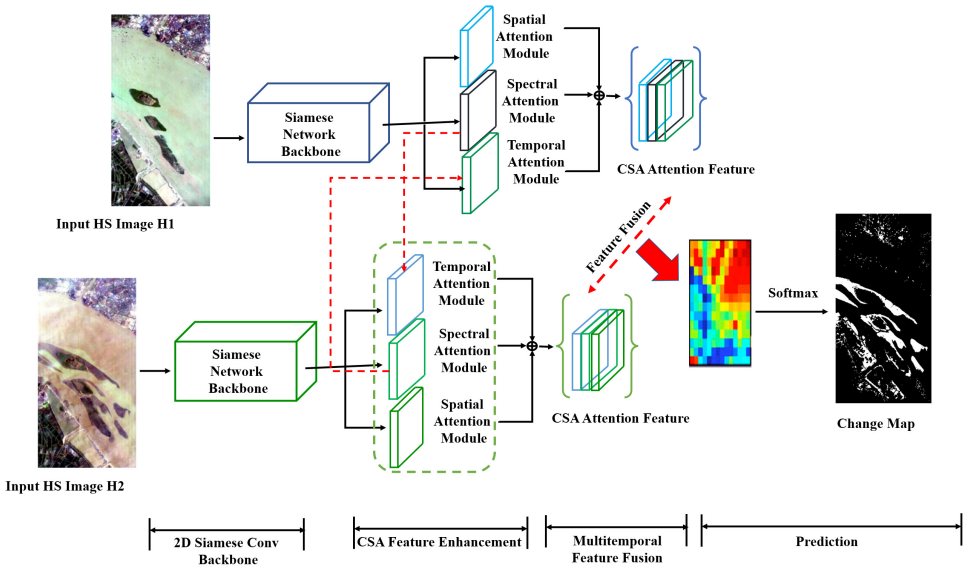

Fig. 1. Overview of the proposed CSA network.

the changes. The major contributions of this letter can be summarized as follows.

1) The novel cross-temporal interaction symmetric attention (CSA) is proposed. In CSA, self-attention is combined to enhance the feature representation ability of each temporal image, and cross-temporal attention is utilized to integrate the difference features oriented from each temporal feature embedding.

2) CSA is introduced into the Siamese 2-D CNN structure to construct the CSANet for HS image CD. CSANet can effectively extract and integrate the joint spatial–spectral–temporal features of the HS images, and at the same time enhance the feature discrimination ability of the changes.

3) CSANet is comprehensively compared with some state-of-the-art (SOTA) deep learning-based HS CD methods, and experimental results demonstrate the effectiveness of the proposed method. The proposed method can yield an average of 1.07% absolute improvement in terms of overall accuracy (OA) and an average of 4.1 absolute improvements in terms of the Kappa coefficient.

II. Proposed Approach

The proposed CSANet is designed by using the attention mechanism to build a symmetric bi-temporal HS image CD framework. The framework of our proposed method is illustrated in Fig. 1. The proposed framework mainly includes three symmetric stages, namely 2-D Siamese Conv backbone, CSA feature enhancement, and multitemporal feature fusion. Initially, the paired HS images are fed into the Siamese 2-D CNN backbone to extract shallow-level features of the HS images. Then, by introducing CSA into the network, joint spatial–spectral–temporal feature integration and enhancement is released. The difference features of the bi-temporal HS images are generated in the last stage to obtain pixel-wise CD results.

A. Cross-Temporal Interaction Symmetric Attention

We propose a novel Cross-Temporal Interaction Symmetric Attention (CSA) module to interact the cross-temporal information of the bi-temporal HS images. The overall structure of the CSA module is shown in Fig. 2. As shown in Fig. 2, the proposed CSA is incorporated with a spatial attention module, a spectral attention module, and a temporal attention module. Specifically, a novel temporal attention module is proposed to interact with the bi-temporal HS image features. Meanwhile, spatial attention and spectral attention are combined to enhance the feature representation ability of each temporal image.

1) Spatial Attention Module: To highlight the spatial information of each temporal image, the spatial self-attention mechanism is introduced to each temporal branch of the network. For the input shallow feature $\mathbf{F}_T \in \mathcal{R}^{H\times W\times C}, T \in (1,2)$ obtained from the upper network of each temporal branch, by applying three convolution operation with 1×1 kernels, the Query, Key, and Value features $\{\mathbf{F}_{T,Q}, \mathbf{F}_{T,K}, \mathbf{F}_{T,V}\} \in \mathcal{R}^{H\times W\times C}, T \in (1,2)$ of $\mathbf{F}_T$ are generated. To obtain the spatial attention matrix $\mathbf{F}_{T,\mathrm{Spa}M}$ of each temporal feature, matrix multiplication, and Softmax operations are adopted between $\mathbf{F}_{T,Q}$ and the transpose of $\mathbf{F}_{T,K}$

$$\mathbf{F}_{T,\mathrm{Spa}M} = \mathrm{softmax}\left(\frac{\mathbf{F}_{T,Q}\mathbf{F}_{T,K}^{\mathrm{Trans}}}{\sqrt{C}}\right), \quad T \in (1,2). \tag{1}$$

The spatial attention feature $\mathbf{F}_{T,\mathrm{Spa}A}$ of each temporal feature $\mathbf{F}_T$ is obtained by

$$\mathbf{F}_{T,\mathrm{Spa}A} = \mathbf{F}_T + \mathbf{F}_{T,\mathrm{Spa}M}\mathbf{F}_{T,V}, \quad T \in (1,2) \tag{2}$$

where C is the dimension of the input data, and $\mathbf{F}_{T,K}^{\mathrm{Trans}}$ is the transpose of $\mathbf{F}_{T,K}$.

2) Spectral Attention Module: The land-cover changes will bring significant variation in the spectral information of the HS images. In this way, enhancing the spectral information

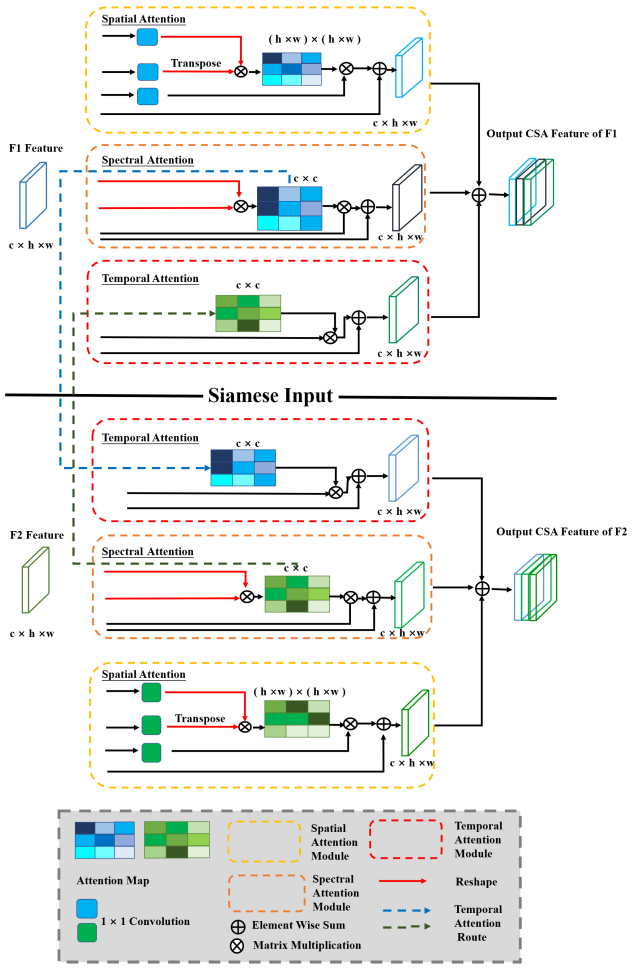

Fig. 2. Overall structure of the cross-temporal interaction symmetric attention.

can effectively improve the network's ability to capture subtle changes. To calculate the spectral attention feature of each input shallow feature, $\mathbf{F}_T$, $\mathbf{F}_T$ are transformed into the $\mathbf{F}_{T,R} \in \mathcal{R}^{M\times C}$, $T \in (1,2)$ space using reshape operation, where $M = H \times W$ represent the shape of $\mathbf{F}_T$. Similarly, the spectral attention matrix is calculated by matrix multiplication and softmax operations between $\mathbf{F}_{T,R}$, $T \in (1,2)$ and its transpose

$$\mathbf{F}_{T,\mathrm{Spe}M} = \mathrm{softmax}\left(\frac{\mathbf{F}_{T,R}\mathbf{F}_{T,R}^{\mathrm{Trans}}}{\sqrt{C_T}}\right), \quad T \in (1,2). \tag{3}$$

The spectral attention $\mathbf{F}_{T,\mathrm{Spe}A}$ of each temporal feature $\mathbf{F}_T$ is obtained by

$$\mathbf{F}_{T,\mathrm{Spe}A} = \mathbf{F}_T + \mathbf{F}_{T,\mathrm{Spe}M}\mathbf{F}_T, \quad T \in (1,2) \tag{4}$$

where C is the dimension of the input data, and $\mathbf{F}_{T,R}^{\mathrm{Trans}}$ is the transpose of $\mathbf{F}_{T,R}$.

3) Temporal Attention Module: To generate the difference feature of the bi-temporal HS images and improve the joint spatial–spectral–temporal feature representation ability of the network, a novel temporal attention mechanism for HS image CD is proposed in this section. Different from spatial attention and spectral attention, which are designed as self-attention, temporal attention can interactively depict the cross-temporal features of the bi-temporal HS images.

The temporal attention is developed based on spectral attention of each temporal feature $\mathbf{F}_T$. To depict the difference feature of the bi-temporal HS images, we introduce the spectral attention matrix obtained by the complementary temporal feature to generate the temporal attention features, thus highlighting the changes. The temporal attention features of $\mathbf{F}_T$ respectively calculated by

$$\mathbf{F}_{1,\mathrm{Temp}A} = \mathbf{F}_1 + \mathbf{F}_{2,\mathrm{Spe}M}\mathbf{F}_1 \tag{5}$$

$$\mathbf{F}_{2,\mathrm{Temp}A} = \mathbf{F}_2 + \mathbf{F}_{1,\mathrm{Spe}M}\mathbf{F}_2. \tag{6}$$

The attention matrix can reflect the relationship between the pixel points. Therefore, for the changed points, the attention matrix of the bi-temporal features presents different characteristics. Information interaction between the temporal features can highlight the changed information.

After obtaining the spatial attention features, spectral attention features and temporal attention features of $\mathbf{F}_T$, the cross-temporal interaction symmetric attention is generated as

$$\mathbf{F}_{T,\mathrm{CSA}} = \left(\mathbf{F}_{T,\mathrm{Spa}A}||\mathbf{F}_{T,\mathrm{Spe}A}||\mathbf{F}_{T,\mathrm{Temp}A}\right), \quad T \in (1,2) \tag{7}$$

where $||$ denotes the concatenation operation.

B. CSANet for HS Image Change Detection

1) Network Structure: By introducing the proposed CSA mechanism, in this section, a novel HS image CD framework is proposed. Assume $H_1 \in \mathcal{R}^{P\times Q\times D}$ and $H_2 \in \mathcal{R}^{P\times Q\times D}$ are two co-registered multitemporal HS images, which were captured in the same geographical location at times T_1 and T_2, respectively. The size of the HS images is $P \times Q \times D$, where D is the number of spectral bands of the co-registered multitemporal HS images. At the same time, $N = P \times Q$ indicates the number of pixels within the analyzing scene. As depicted in Fig. 1, two temporal branches $\mathcal{G}_{T_1}$ and $\mathcal{G}_{T_2}$ are designed to simultaneously extract the joint spatial–spectral–temporal features of the original bi-temporal HS image data. For each temporal branch $\mathcal{G}_T$, the HS image patches are first fed into the 2D-CNN Blocks to obtain the shallow-level features $\mathbf{F}_T$ of each temporal image. First, 2D-Convolution with a convolutional kernel of 3×3 is adopted to extract the semantic information of the HS images. For each block, only 2-D Conv, 2-D BN, Relu, and MaxPooling operations are included.

After obtaining the shallow-level features $\mathbf{F}_T$ of each temporal image, CSA mechanism is applied to generate the attentional features $\mathbf{F}_{T,\mathrm{CSA}}$ of each temporal. To generate the difference features of the HS images, before fed into the fully connected layers, concatenation operation is adopted between $\mathbf{F}_{T,\mathrm{CSA}}$. The output of the fully connected layers can be depicted as

$$\mathrm{FC}\left(\mathbf{F}_{1,\mathrm{CSA}}, \mathbf{F}_{2,\mathrm{CSA}}\right) = f\left(W \cdot \left(\mathbf{F}_{1,\mathrm{CSA}}||\mathbf{F}_{2,\mathrm{CSA}}^{L}\right) + b\right) \tag{8}$$

where W and b, respectively, represent the weight and bias of the fully connected layer. Then, the HS difference feature is fed into the Softmax Layer to predict the changes.

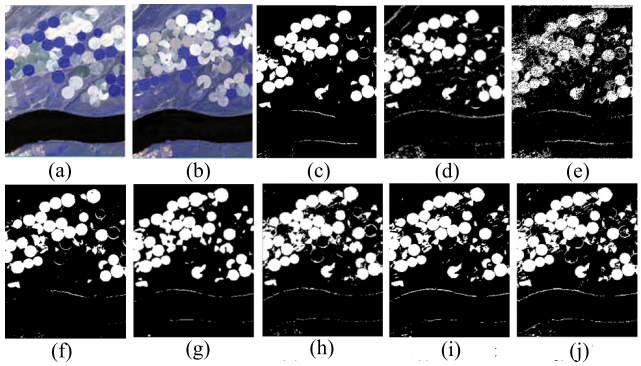

Fig. 3. Testing bi-temporal HS images and binary change maps generated by methods on the irrigated agricultural dataset. (a) Before. (b) After. (c) CVA. (d) DSFA. (e) PBCNN. (f) GETNET. (g) DBDA. (h) SSCNN-S. (i) CSANet. (j) GT.

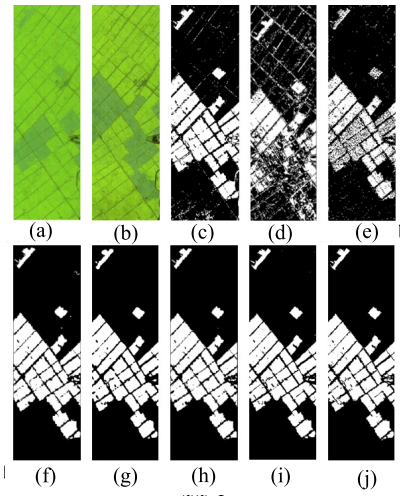

Fig. 4. Testing bi-temporal HS images and binary change maps generated by methods on the wetland agricultural dataset. (a) Before. (b) After. (c) CVA. (d) DSFA. (e) PBCNN. (f) GETNET. (g) DBDA. (h) SSCNN-S. (i) CSANet. (j) GT.

2) Training Strategies: In the process of network training, the training set samples are input into the proposed framework in the form of $\{\mathbf{P}_{T_1,i}, \mathbf{P}_{T_2,i}, y_i\}_{i=1}^{\mathrm{NT}}$, where $\mathbf{P}_{T,i}$ are the HS image training patches, and $y_i \in (0, 1)$ is the real labels corresponding to the pairs. NT is the total number of training samples. As class imbalance is a common phenomenon in the HS image CD task, the cross-entropy is adopted as the loss function

$$\mathcal{L}_{W-C} = -\frac{1}{\mathrm{NT}} \sum \sum \omega_j \log \Pr(y_i = j | p_i, \theta) \tag{9}$$

where ω_j is the jth weight and $\Pr(y_i = j|p_i, \theta)$ is the probability that the pixel belongs to the jth class.

III. Experiments

To demonstrate the effectiveness of the proposed HS image CD framework, a series of experiments have been conducted on three real-world bi-temporal HS images.

A. Dataset Description

In this letter, three public real-world bi-temporal HS CD datasets are used as testing datasets. The three involved datasets were all captured by the Earth Observing-1 (EO-1) Hyperion sensor. The first dataset is the Irrigated Agricultural Dataset, which depicts an irrigated agricultural area of Hermiston city in Umatilla County Oregon, USA [8]. The second dataset is the Wetland Agricultural Dataset, which depicts a farmland area of Yuncheng City, Zhejiang Province, China [8]. The third dataset is the River dataset [4], which depicts a river area in Jiangsu Province, China. Each dataset is divided into three subsets including the training set, validation set, and testing set. Table I gives the detailed information of the three datasets.

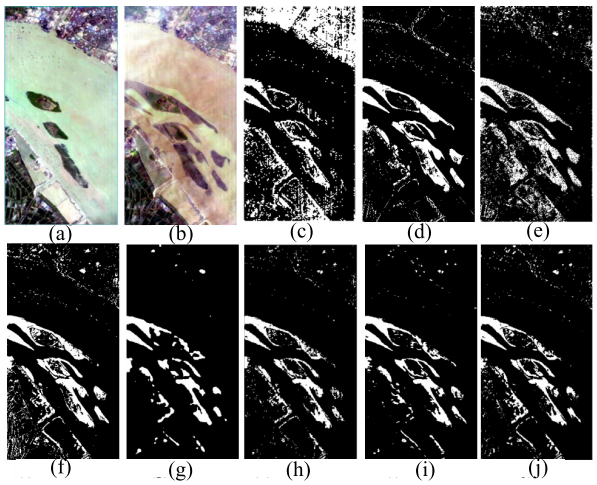

Fig. 5. Testing bi-temporal HS images and binary change maps generated by methods on the river dataset. (a) Before. (b) After. (c) CVA. (d) DSFA. (e) PBCNN. (f) GETNET. (g) DBDA. (h) SSCNN-S. (i) CSANet. (j) GT.

TABLE I
DATASETS INFORMATION

Dataset	Irrigated Agricultural Dataset	Wetland Agricultural Dataset	River Dataset
Size	307×241	450×140	463×241
Spectral Bands	156	156	198
Date T_1	May 1st, 2004	May 3rd, 2006	May 3rd, 2013
Date T_2	May 8th, 2007	Apr 23rd, 2007	Dec 31st, 2013
Training Percentage	9.77%	20.95%	3.36%
Training Size	7232	13200	3750

B. Results and Discussions

To demonstrate the effectiveness of our proposed method, we compare our proposed method with some recently proposed CD methods, including CVA [9], DSFA [10], PBCNN [11], GETNET [4], DBDA [12], and SSCNN-S [5], where CVA and DSFA are designed in an unsupervised way. The performance of each comparison method is assessed by several accuracy evaluation indexes including overall accuracy (OA) and kappa coefficient (K). The learning rate is set as 0.0005. Meanwhile, the input patch size of the proposed methods is set as 9×9, which can achieve a balance between classification accuracy and efficiency. To make a fair comparison, in the experiments, we use k-means clustering as the automatic threshold selection method for the unsupervised CD methods. For the deep learning-based methods, we adopt the same parameter settings introduced in the corresponding

TABLE II
QUANTITATIVE CHANGE DETECTION EVALUATIONS OF THE DIFFERENT DATASETS

Dataset	Irrigated Agricultural Dataset		Wetland Agricultural Dataset		River Dataset	
Measure	OA	Kappa	OA	Kappa	OA	Kappa
CVA	0.931	0.782	0.790	0.499	0.812	0.337
DSFA	0.949	0.834	0.783	0.673	0.942	0.727
PBCNN	0.890	0.669	0.918	0.794	0.913	0.890
GETNET	0.943	0.825	0.975	0.939	0.949	0.747
DBDA	0.963	0.892	0.978	0.947	0.959	0.726
SSCNN-S	0.965	0.891	0.977	0.944	0.964	0.743
CSANet	**0.980**	**0.941**	**0.990**	**0.975**	**0.968**	**0.785**
Time	Training	Testing	Training	Testing	Training	Testing
(s)	5.54	16.84	9.81	17.38	3.30	27.21

TABLE III
ABLATION STUDIES ON ATTENTION MODULES FOR DIFFERENT DATASETS

Dataset			Irrigated Agricultural Dataset		Wetland Agricultural Dataset		River Dataset	
SpaA	**SpeA**	**TemA**	OA	Kappa	OA	Kappa	OA	Kappa
√			0.938	0.861	0.943	0.919	0.932	0.715
	√		0.942	0.865	0.951	0.927	0.931	0.718
√	√		0.957	0.886	0.969	0.940	0.951	0.726
√	√	√	**0.980**	**0.941**	**0.990**	**0.975**	**0.968**	**0.785**

papers. For the implementation of our program, all the experiments are implemented using Python 3.6, the convolutional neural networks are constructed with PyTorch. Experiments are conducted using a personal computer equipped with 6 GB of memory and an NVDIA GeForce RTX 1660.

Table II provides the quantitative CD performance results, training time(/epoch), and testing time for the three involved testing datasets. From the results, the proposed method outperforms all the SOTA deep learning-based HS image CD comparison methods, and a significant improvement can be observed both in terms of OA and Kappa. For instance, for the Irrigated Agricultural Dataset, the proposed CSANet achieved an OA of 98.0%, which yields an absolute improvement of 2.5% compared to SSCNN-S, which demonstrates the joint spatial–spectral–temporal feature extraction ability of the CSANetwork.

Meanwhile, Figs. 3–5, respectively, provides the visualization results of the SOTA methods on the testing datasets. From the figures, the proposed method produces a better CD performance. Specifically, compared with DBDA, which is also developed based on the attentional mechanism, the proposed method produces a more accurate change map with clear boundaries and less noise, which demonstrates the effectiveness of the temporal attention in depicting the difference feature between the temporal images.

C. Ablation Study

The attentional modules applied in the CSA play a significant role in the CD accuracy. In the proposed framework, we combine spatial attention, spectral attention, and temporal attention to generate CSA. To demonstrate the effectiveness of the adopted attentional modules, we fix all the other parameters to compare the CD results when varying the adopted attentional modules. Table III shows the HS CD performance on the three datasets. From the results, significant improvement can be observed in the application of temporal attention, which demonstrates the multidimensional feature extraction ability of CSA.

IV. CONCLUSION

In this letter, a novel Cross-Temporal Interaction Symmetric network is proposed for HS image CD. In CSA, self-attention is combined to enhance the feature representation ability of each temporal image, and cross-temporal attention is utilized to integrate the difference features oriented from each temporal feature embedding. On this basis, an HS image CD framework based on CSA and Siamese 2-D CNN is proposed for the joint spatial–spectral–temporal features integration of the bi-temporal HS images. Experimental results on three real-world HS image CD datasets demonstrate the superiority of our proposed CD framework when compared with SOTA deep learning-based CD methods.

The proposed CSA module provides reference significance for spatial–spectral–temporal feature fusion in different research fields. In future work, we will continue to explore a more general CD framework to deal with HS image CD tasks under complex conditions.

REFERENCES

[1] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[2] 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.

[3] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[4] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2018.

[5] T. Zhan *et al.*, "SSCNN-S: A spectral-spatial convolution neural network with Siamese architecture for change detection," *Remote Sens.*, vol. 13, no. 5, pp. 895–907, 2021.

[6] 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, 2022, Art. no. 5507705.

[7] D. He, Q. Shi, X. Liu, Y. Zhong, and L. Zhang, "Generating 2m fine-scale urban tree cover product over 34 metropolises in China based on deep context-aware sub-pixel mapping network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, Feb. 2022, Art. no. 102667.

[8] M. Hasanlou and S. Seydi, "Hyperspectral change detection: An experimental comparative study," *Int. J. Remote Sens.*, vol. 39, no. 20, pp. 7029–7083, 2018.

[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] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[11] A. Sharma, X. Liu, X. Yang, and D. Shi, "A patch-based convolutional neural network for remote sensing image classification," *Neural Netw.*, vol. 95, pp. 19–28, Nov. 2017.

[12] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, pp. 582–606, 2020.