CBANet: An End-to-End Cross-Band 2-D Attention Network for Hyperspectral Change Detection in Remote Sensing

Yinhe Li, Jinchang Ren, *Senior Member, IEEE*, Yijun Yan, *Member, IEEE*, Qiaoyuan Liu, Ping Ma, Andrei Petrovski, and Haijiang Sun

***Abstract*— As a fundamental task in remote sensing (RS) observation of the earth, change detection (CD) using hyperspectral images (HSI) features high accuracy due to the combination of the rich spectral and spatial information, especially for identifying land-cover variations in bi-temporal HSIs. Relying on the image difference, existing HSI CD methods fail to preserve the spectral characteristics and suffer from high data dimensionality, making them extremely challenging to deal with changing areas of various sizes. To tackle these challenges, we propose a cross-band 2-D self-attention network (CBANet) for end-to-end HSI CD. By embedding a cross-band feature extraction module into a 2-D spatial–spectral self-attention module, CBANet is highly capable of extracting the spectral difference of matching pixels by considering the correlation between adjacent pixels. The CBANet has shown three key advantages: 1) less parameters and high efficiency; 2) high efficacy of extracting representative spectral information from bi-temporal images; and 3) high stability and accuracy for identifying both sparse sporadic changing pixels and large changing areas whilst preserving the edges. Comprehensive experiments on three publicly available datasets have fully validated the efficacy and efficiency of the proposed methodology.**

***Index Terms*— Change detection (CD), cross-band self-attention network, hyperspectral images (HSI), spatial–spectral feature extraction.**

I. Introduction

CHANGE detection (CD) task can identify differences in multi-temporal remote sensing (RS) imageries within the same geographic area [1]. In recent years, hyperspectral images (HSI) have been successfully applied for RS observation of the earth [2]. With the 2-D spatial information and rich spectral information in the third dimension, HSI can acquire continuous narrow bands with a high spectral resolution [3]. Compared with multi-spectral images and conventional color images in red-green-blue (RGB), HSI has the following two advantages: 1) high spectral resolution and wide spectral range spanning from visible light to short-wave infrared, even mid-infrared, where the spectral resolution can be 10 nm or even less along with hundreds of continuous bands [4] and 2) rich spatial and spectral information for effective detection of the region of interests [5]. Therefore, hyperspectral CD (HCD) has become a research hotspot, which has been successfully applied in a wide range of applications such as precision agriculture [6], disaster monitoring [7], geological survey [8], and biomedical science [9], etc.

However, there are still challenges summarized as follows:

1) Most existing CD methods rely on the difference between the bi-temporal hypercubes, in which the spectral characteristics can be damaged [10].
2) Existing deep learning (DL) models for HCD have a large amount of hyperparameters, resulting in redundant information in both spatial and spectral domains as well as a large computational cost [11].
3) Most of the HCD methods fail to deal with sparsely distributed changing areas of various sizes [12].

To tackle these issues, a lightweight DL network, namely cross-band 2-D self-attention network (CBANet), is proposed, which fuses the cross-band module for extracting spectral domain features pixel-by-pixel and designs a new 2-D self-attention module [13] for improved extraction of local spatial–spectral features whilst keeping the network compact for efficiency. The major contributions are summarized as follows:

1) A cross-band feature extraction module is proposed to extract the mutual and representative features from bi-temporal hypercubes, where a 1 × 1 convolutional layer is introduced to greatly increase the non-linear characteristics (using the subsequent activation function) of the feature map while keeping the scale of the feature map unchanged.
2) A 2-D self-attention module is proposed for focused extraction of local spatial–spectral features and improved feature representation and discrimination capability, resulting in enhanced network reliability.

Manuscript received 29 November 2022; revised 2 March 2023 and 25 April 2023; accepted 11 May 2023. Date of publication 16 May 2023; date of current version 6 June 2023. This work was supported in part by the International Cooperation Project of Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP) under Grant Y9U933T190; in part by the Key Discipline Improvement Project of Guangdong Province under Grant 2022ZDJS015; and in part by the Dazhi Scholarship of GPNU. *(Corresponding authors: Jinchang Ren; Yijun Yan; Qiaoyuan Liu.)*

Yinhe Li, Yijun Yan, Ping Ma, and Andrei Petrovski are with the National Subsea Centre, Robert Gordon University, AB21 0BH Aberdeen, U.K. (e-mail: y.yan2@rgu.ac.uk).

Jinchang Ren is with the National Subsea Centre, Robert Gordon University, AB10 7AQ Aberdeen, U.K., and also with the School of Computer Science, Guangdong Polytechnic Normal University (GPNU), Guangzhou 510651, China (e-mail: jinchang.ren@ieee.org).

Qiaoyuan Liu and Haijiang Sun are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China (e-mail: liuqy@ciomp.ac.cn).

Digital Object Identifier 10.1109/TGRS.2023.3276589

3) A novel end-to-end lightweight CBANet is proposed which can produce higher detection accuracy but has fewer hyperparameters. Its efficacy and efficiency have been fully validated in comprehensive experiments when compared with a few state-of-the-art approaches.

The remainder of this article is organized as follows. Section II introduces the related work for HCD. Section III describes the details of the proposed CBANet. Section IV presents the experimental results and assessments. Finally, some remarkable conclusions are summarized in Section V.

II. Related Work

In the last decades, numerous supervised and unsupervised algorithms have been developed for HCD tasks. Generally, these algorithms can be categorized into the following three categories, i.e., algebraic operation-based, image transformation-based, and DL-based methods. Some early studies on HCD focus mainly on algebraic operations. In [14], the changing regions are first identified by a pixel-wise difference based on the absolute distance (AD). Change vector analysis (CVA) [15] was proposed to compare the magnitude of each vector pair and calculate the Euclidean distance. Spectral angle mapping (SAM) [16] was used to determine the spectral similarity of corresponding pixels from dual-time HSI for CD. The final decision will be made by thresholding of the changes or clustering. As seen, the algebraic operation-based methods are straightforward and have a relatively low computation cost. However, the redundant information caused by the high correlation between bands brings a barrier to precise HSI CD.

To solve these issues, image transformation-based methods have been widely explored in the last two decades. Typical methods include the principal component analysis (PCA) [17], independent component analysis (ICA) [18], and linear discriminant analysis (LDA) [19], in which the image transformation-based methods can help to convert the high-dimensional spectral data into a new low-dimensional representation whilst retaining the discriminative information and reducing the data redundancy. However, the accuracy of retaining key information is susceptible to an unbalanced distribution of the data of the distorted data statistics. To tackle this issue, multivariate alteration detection (MAD) [20] is developed on the basis of the canonical correlation analysis (CCA) [21] to seek the linear combination of the orthogonal differences between the pair of corresponding bands from dual-time images to determine the maximum variance between samples, dividing the features by the statistic method of the chi-square distribution. Iteratively reweight (IR) MAD [22] is presented to assign each changing feature a weight, which is calculated by the sum of squares of the standardized MAD to measure the degree of change, resulting in a more accurate and less noisy binary change map than the MAD, that the sum of squared standardized MAD variates is small, large weights refer to little change and vice versa. Although image transformation-based methods can effectively remove the data redundancy, it ignores the spectral continuity and damages the similarity between adjacent pixels when mapping the original image into the new low-dimensional spectral domain [23].

Recently, it became a new trend to apply DL-based methods to HCD tasks for extracting more effective and representative spectral, spatial, and spectral–spatial features. In [24], a generic end-to-end 2-D convolutional neural network (CNN) is introduced, using a mixed affinity matrix with subpixel representation to mine cross-band gradient. In [25], a recurrent 3-D fully convolutional network is proposed, in which 3-D CNN layers are employed to extract spectral–spatial features whilst multi-temporal change features are extracted by combining CNN and the long-short-term-memory (LSTM) model. In [26], a multilevel encoder-decoder attention network is proposed to extract more effective hierarchical spatial–spectral features, where the encoder-decoder module transfers the feature to the LSTM for analyzing the temporal dependence. In [27], a CNN framework with slow-fast band selection and feature fusion grouping was proposed to extract changed features. In [28], a novel noise modeling-based unsupervised fully convolutional network is proposed for improved extraction of the discriminative CNN features. Although DL-based models produce quite good results, they often rely on a large volume of training data, which can be unavailable in real cases, and very high computational cost thus needs to be further addressed [29].

The self-attention modules are also widely adopted in DL-based models for improved feature extraction and enhanced accuracy and robustness of classification. In [30], a novel cross-temporal interaction symmetric attention (CSA) network was proposed, where a Siamese network was equipped to hierarchically extract the change information in a symmetric pattern. By extracting the joint spatial–spectral–temporal features of the bi-temporal images, a cross-temporal self-attention module was combined to integrate the difference maps from each temporal feature embedding and enhance the feature representation ability for a more accurate and reliable CD. In [31], a deep multi-scale pyramid network was proposed that aggregated the multi-scale features, level by level, where a spatial–spectral residual attention module was applied to further enhance the features by making the network pay more attention to significant information. In [32], a pixel-level self-supervised hyperspectral spatial–spectral feature understanding network was proposed for pixel-wise feature representation instead of 2-D band-based processing, where a powerful spatial–spectral attention module based on fully convolutional layers was employed to explore the spatial correlation and discriminative spectral features. In [33], a joint spectral, spatial, and temporal transformer network (SST-Former) was proposed, using multi-head self-attention modules as the input to improve the feature extraction, where various encoders and decoders were employed to extract the sequence information in the spectral–spatial domain before determining the changes via a residual structure based cross attention module.

III. Methodology

The diagram of the proposed CBANet is presented in Fig. 1, which is composed of main three modules: 1) cross-band spectral feature extraction; 2) spectral–spatial feature extraction, and 3) 2-D self-attention-based deep feature extraction. The specific details of the network are shown in Table I.

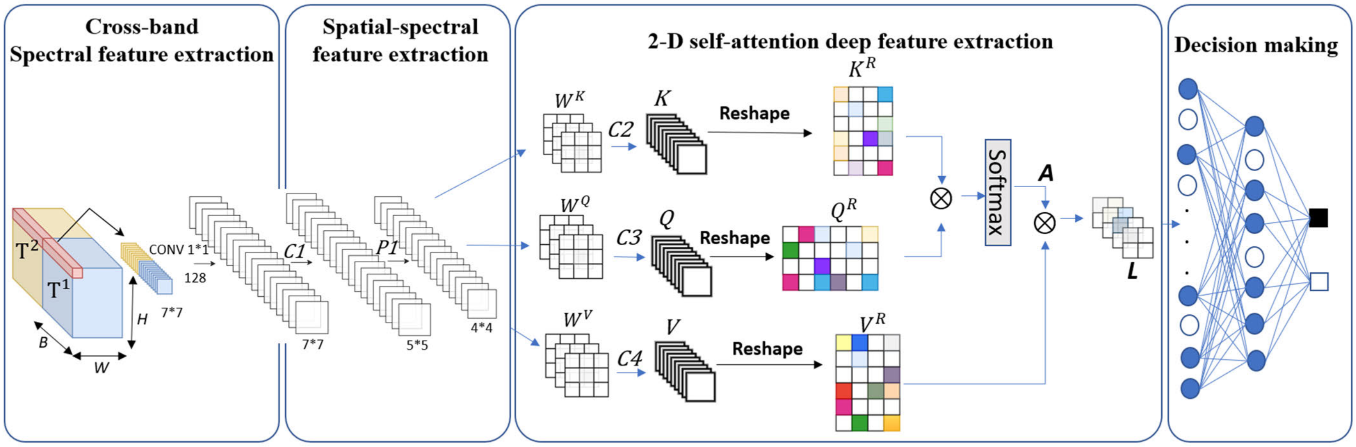


Fig. 1. Architecture of the proposed CBANet model.

TABLE I
ARCHITECTURE DETAILS FOR PROPOSED MODEL, WHERE B IS THE NUMBER OF BANDS

Layers	Type	No. Kernel	Size
Input	-	B*2	-
Conv1	Conv2-D+BN	128	1×1
$C1$	Conv2-D+BN	128	3×3
$P1$	Average Pooling	-	2*2
$C2$	Conv2-D+BN	32	3×3
$C3$	Conv2-D+BN	32	3×3
$C4$	Conv2-D+BN	32	3×3
Flatten	Flatten	512	-
$FC1$	Linear (Dropout=0.4)	64	-
$FC2$	Linear (Dropout=0.4)	16	-
$FC3$	Linear (Dropout=0.4)	2	-

A. Cross-Band Spectral Feature Extraction

Given a pair of spatially aligned bi-temporal hypercubes $T^1 \in \mathcal{R}^{W*H*B}$ and $T^2 \in \mathcal{R}^{W*H*B}$, where W and H denote the width and height of the spatial size, and B represents the number of spectral bands. T^1 and T^2 are first concatenated to form a new hypercube $Q \in \mathcal{R}^{W*H*2B}$, which will be divided into a group of overlapped 3-D neighboring patches denoted as $Z_{(\alpha,\beta)} \in \mathcal{R}^{S*S*2B}$, where S is the spatial size of Z, (α, β) denote the coordinates of the patch center in the spatial domain where $\alpha\epsilon[1, W]$, $\beta\epsilon[1, H]$. The total number of 3-D patches from Q will be $(W - S + 1) \times (H - S + 1)$. For each patch $Z_{(.)}$, the whole spectral vector may contain highly redundant information and cause huge computational costs in training the DL model. Thus, reducing the data dimension whilst keeping the discriminative information in the spectral domain becomes the key issue here. For this purpose, a 1 × 1 convolutional layer [13] with a proper setting of k_{Conv1} kernel's number is applied to the dual spectral bands $Z_{(.)}$, the weighted fusion across the whole spectral vector can help to compose a new feature fusion space with much lower spectral dimensionality. Meanwhile, the input patch size S of the proposed methods is set to 7 × 7 and the number of kernels k_{Conv1} in the cross-band feature extraction module is set to 128, as it can achieve a good balance between the computational efficiency and the retained principal components.

B. Spectral–Spatial Feature Extraction

The low-dimensional feature cube is constructed after extracting the spectral features by passing through the cross-band fusion module, which is a 1 × 1 convolutional layer to preserve the characteristics of the bi-temporal cubes and remove redundant information. In the next step, a 2-D convolutional kernel is employed for global feature extraction in the spatial domain. The convolution sums up the dot product between the input feature map and the kernel. The 2-D kernels are stride over the input feature map to cover the entire spatial domain. The convolutional results with an adding on the additional bias will pass through a ReLu function. In 2-D convolution, the jth feature map in the ith layer at position (x, y) is denoted as $F_{i,j}^{x,y}$, which is calculated as follows:

$$F_{i,j}^{x,y} = \text{ReLu}\left(\text{BN}\left(b_{i,j} + \sum_{r}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{i,j,r}^{p,q} F_{i-1,r}^{x+p,y+p}\right)\right) \tag{1}$$

where P_i, Q_i are the height and width of the 2-D kernel, $b_{i,j}$ is the bias, $w_{i,j,r}^{p,q}$ is the weight parameter at the position (p, q) of the kernel connected to the rth feature map, where r represents the set of feature maps in the $(i-1)$th layer connected to the ith layer [34]. ReLu$(\cdot)$ is the rectified linear unit [35] as an activation function to introduce the nonlinearity, reduce parameter interdependence and alleviate overfitting BN$(\cdot)$ represents the batch normalization function. In this module, 2-D convolution with a kernel size of 3 × 3 is used in order to reduce the network parameters as well as extract more representative local information. Afterward, 2 × 2 sub-sampling average pooling is adopted to prevent features from rotation and scale during convolution [36]. The extracted spectral–spatial features are represented as $X \in \mathcal{R}^{h*h*k_{C1}}$, where $h = 4$ after pooling. The number of kernels k_{C1} for spectral–spatial feature extraction is set to 128, as it

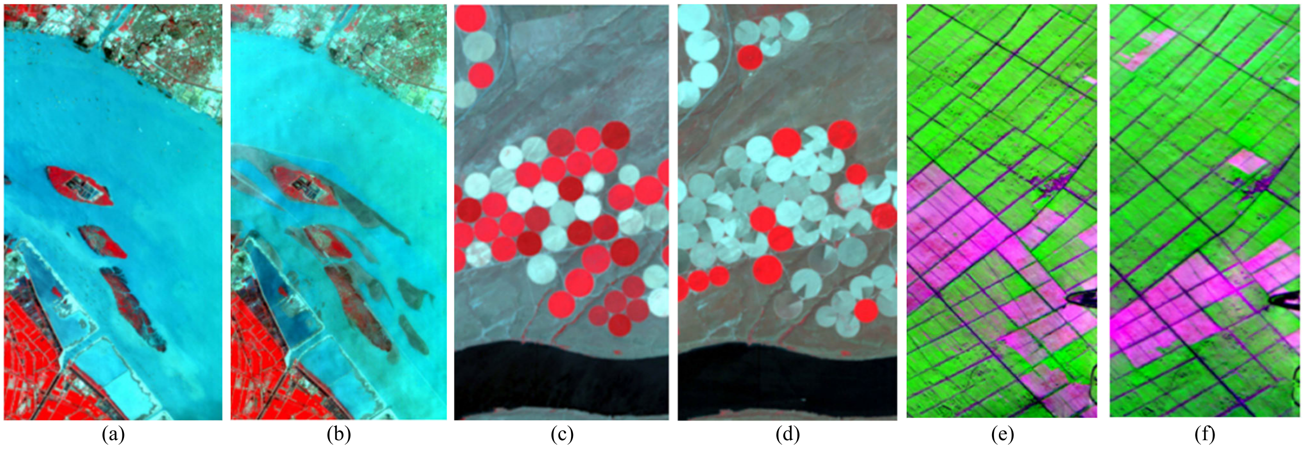


Fig. 2. Pseudo-color images of the three datasets. (a) River on May 3, 2013. (b) River on December 31, 2013. (c) Hermiston on May 1, 2004. (d) Hermiston on May 8, 2007. (e) Yancheng on May 3, 2006. (f) Yancheng on April 23, 2007.

reaches a good tradeoff between classification accuracy and robustness.

C. 2-D Self-Attention Based Deep Feature Extraction

Previous studies have found that the self-attention mechanism has beneficial to conventional CD tasks [37] and HSI classification [38], [39]. However, these self-attention models use the 1 × 1 convolutional kernel and focus on pixel-wise band features to assign the pixel-wise weights and only pay attention to the spectral information, leading to insufficient detection performance especially when dealing with the changing areas in various sizes. Motivated by this issue and inspired by the work in [40], we propose a 2-D self-attention module to build adjacent pixels dependency in local space as well as enhance the spatial–spectral features from the middle level toward the deeper level. The features X is taken as the input and fed into three 3×3 2-D convolutional layers ($C2$, $C3$, $C4$) to generate three new spatial feature maps, denoted as Query (Q), Key (K), and Value (V), where $(K, Q, V) \in \mathcal{R}^{m*m*k_{C2}}$, we set $k_{C2} = k_{C3} = k_{C4} = 32$ in this study. Each feature map will be converted to 2-D attention matrices denoted as K^R, Q^R, V^R, respectively, where $(K^R, V^R) \in \mathcal{R}^{m^2*k_{C2}}$ and $Q^R \in \mathcal{R}^{k_{C2}*m^2}$. Then the correlation can be obtained by the dot product of the attention matric K^R and Q^R, from the properties of the dot product, the higher similarity between the two matrics, the value of the dot product will be larger that represents the more obvious local change feature, and will be assigned a greater weight. The spatial attention matrix A is calculated by multiplication between K^R and Q^R followed by Softmax operations:

$$A = \text{Softmax}\left(K^R * Q^R\right). \tag{2}$$

Finally, the 2-D attention feature map $L \in \mathcal{R}^{m*m*k_{C2}}$ can be obtained by multiplying A by V^R. In this process, all local features are involved in the calculation, therefore, 2-D self-attention not only can capture the global feature distribution, but also focus on the key changing features. The larger the weight value in the feature map L, the more prominent the feature.

Since CD can be considered as a binary classification problem of distinguishing the change and non-change pixels, the cross entropy, which is commonly used for classification, is adopted as the loss function

$$\text{Loss}_{(\text{pred},\text{label})} = -\frac{1}{u}\sum_{i=1}^{n}(l * \log(p) + (1-l) * \log(1-p)) \tag{3}$$

where u denotes the number of samples, l represents the ground truth value, and 0 and 1 represent unchanged and changed regions. p represents the probability predicted by the Linear function. The selected optimizer is the adaptive momentum (Adam) [41] with an initial learning rate of 0.0001.

IV. Experiments

A. Dataset Description

All the three datasets we used in the experiments were acquired by the Hyperion sensor mounted onboard the Earth Observing-1 (EO-1) satellite, which offers a total of 242 bands ranging from 0.4–2.5 μm, with a spatial resolution of 10 m and a spectral resolution of 30 nm [42], [43].

The River dataset, shown in Fig. 2(a) and (b), was collected over the Jiangsu Province, China on May 3, 2013, and December 31, 2013 [27], respectively. After noise removal and image registration, this dataset contains 463 × 241 spatial pixels and 198 spectral bands, where the major changed regions are the substance in the river and the structure of the riverbank.

The Hermiston dataset, shown in Fig. 2(c) and (d), was collected in Hermiston city, Oregon, United States on May 1, 2004, and May 8, 2007 [44], respectively. After noise removal and image registration, this dataset contains 390 × 200 spatial pixels and 242 spectral bands, where the changing factors are crop growth situation and the water content of crops that were affected by irrigation conditions in the farmland.

The Yancheng dataset, shown in Fig. 2(e) and (f), was collected in Yancheng city, China on May 3, 2006, and April 23, 2007 [28], respectively. After noise removal and image registration, this dataset contains 420 × 140 spatial pixels and

154 spectral bands, where the major change is the land cover on wetlands.

B. Evaluation Criteria

CD task can be considered as a binary classification problem where the changed pixels and unchanged pixels are presented as 1 and 0, respectively on the extracted binary change map. To quantitatively assess the performance, several commonly used metrics were adopted, which include the overall accuracy (OA), Kappa coefficient (KP), changed cluster detection accuracy (CA) and non-CA (NCA), and average accuracy (AA) that represents the average value of CA and NCA [45]. The OA is the percentage of correctly classified pixels, defined as

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

where TP, TN, FP, and FN denote respectively the correctly detected changed pixels, correctly detected unchanged pixels, incorrectly detected changed pixels, and incorrectly detected unchanged pixels.

KP is to measure the interrater reliability as the degree of similarity between the change map and the ground truth

$$\mathrm{KP} = \frac{\mathrm{OA}-\mathrm{PRE}}{1-\mathrm{PRE}} \tag{5}$$

$$\mathrm{PRE} = \frac{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2} + \frac{(\mathrm{FN}+\mathrm{TN})(\mathrm{FP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2}. \tag{6}$$

For a more intuitive comparison, CA and NCA are used to represent the detection accuracy of the changed cluster and the non-changed cluster, respectively, as given below

$$\mathrm{CA} = \frac{M_C}{N_1}, \mathrm{NCA} = \frac{M_G}{N_0} \tag{7}$$

whereM_C and M_G denote the number of the corrected detected changed and non-changed pixels in the change map, respectively; N_1 and N_0 denote the number of changed and non-changed pixels in the ground truth, respectively.

C. Results and Comparison

In this session, we evaluate the effectiveness of the proposed method by comparing it with a few start-of-the-art unsupervised methods, which include the CVA [22], PCA-KM [24], and AD [21] as well as several DL-based methods such as 2-D-CNN [46], 3-D-CNN [47], HybridSN [35] CSANet [31] and LSTM [48]. It is worth noting that the compared methods except CSANet will need to take the difference of the given HSI pairs as input, which may thus break the continuity of the spectral features. Thanks to the cross-band fusion module used; such image differencing is not needed for our proposed end-to-end network. The proposed CBANet and all other DL-based methods are trained based on the PyTorch on an NVIDIA RTX A2000, with the batch size set to 32 and the number of training epochs as 500. We randomly select 20% of pixels in the changed and unchanged pixels as the training set and the remaining for testing. To make a fairer and more reliable comparison, all DL algorithms are repeated ten times in each experiment, and the averaged results with the standard deviations are reported. In the produced change maps, false alarms, and missing pixels are marked in red and green respectively for ease of comparison, white areas represent correctly detected and black area for true negatives.

1) Experiments on the River Dataset: The experimental results from the River dataset are shown in Fig. 3 and Table II. As seen in the ground-truth map in Fig. 3(j), the most obvious differences are the differently shaped sediment accumulations in the river and the land-cover changes on the riverbank, in addition to many others. In Fig. 3(a)–(c), most of the non-changed pixels are detected as false alarms, distributed in the upper and lower left corners of the maps, and are wrongly detected as changed areas by all unsupervised algorithms. However, most false alarms can be correctly classified by DL-based algorithms. In the regions in the upper left corner of the maps, although most changing pixels can be distinguished by the 3-D-CNN and 2-D-CNN in Fig. 3(f) and (g), some sporadic changing pixels are still not identified, due possibly to that both these DL-based models only extract the relationship between local and global spatial–spectral features but ignoring the changing features of the independent pixels in the spatial domain. The CSANet has produced the second highest OA, CA, AA, and KP values among all compared DL-based models, only slightly worse than our CBANet, owing to the joint spatial–spectral–temporal features extracted by the introduced self-attention module. Also, our CBANet has a much higher CA than the CSANet in effective detection of the changed pixels whilst maintaining the same or even slightly lower level of false alarms as measured by NCA. Thanks to our cross-band fusion module and the 2-D self-attention module, both sporadic changing pixels and large regions can be accurately detected. As shown in Table II, not surprisingly, the DL-based supervised methods all have higher OA and KP and outperform the unsupervised ones. As for CA and AA, however, DL methods seem inferior to unsupervised ones, due mainly to the fact that the changing pixels have the characteristics of wide distribution and various scales. Note that CVA, AD, and PCA-KM are all pixel-wise methods, which do not consider the influence of adjacent pixels and thus are more sensitive to noise. Therefore, they tend to misclassify the changing pixels, resulting in a low NCA. On the contrary, DL algorithms are more accurate in distinguishing small changes. In the benchmarked DL methods, LSTM has the worst performance with an average KP of 0.7261% and OA of 95.69%. Our proposed CBANet has produced the highest OA, KP, and NCA over all compared methods, achieving the highest CA value over all DL methods, which indicates the correct detection of changing areas in various sizes.

2) Experiments on the Yancheng Dataset: According to the HCD results shown in Fig. 4 and Table III, the primary changing area in this dataset is land cover on wetlands, see Fig. 4(j). Again, all three unsupervised methods have quite poor results as shown in Fig. 4(a)–(c), where many changing pixels are missed along with false alarms in striped lines and other occasions, resulting in low values of KP at around 0.71% and OA less than 90% (Table III). Obviously, all the DL-based methods have outperformed the unsupervised

TABLE II
QUANTITATIVE ASSESSMENT OF DIFFERENT METHODS ON THE RIVER DATASET

	OA	CA	NCA	AA	KP
AD	0.9431	0.9423	0.9515	0.9469	0.7137
CVA	0.9253	0.9217	0.9635	0.9425	0.6528
PCA-KM	0.9517	0.9518	0.9505	0.9511	0.7476
LSTM	0.9569±0.0011	0.7671±0.0074	0.9746±0.0019	0.8704±0.0038	0.7216±0.0070
HybridSN	0.9671±0.0019	0.7605±0.0298	0.9867±0.0043	0.8736±0.0130	0.7826±0.0087
3-D-CNN	0.9700±0.0008	0.7888±0.0299	**0.9871**±0.0036	0.8879±0.0124	0.8045±0.0053
2-D-CNN	0.9682±0.0007	0.8346±0.0118	0.9806±0.0021	0.9083±0.0063	0.7946±0.0033
CSANet	0.9743±0.0012	0.8623±0.0049	0.9847±0.0009	0.9235±0.0094	0.8360±0.0049
CBANet	**0.9765**±0.0012	**0.8800**±0.0110	0.9865±0.0018	**0.9308**±0.0065	**0.8526**±0.0036

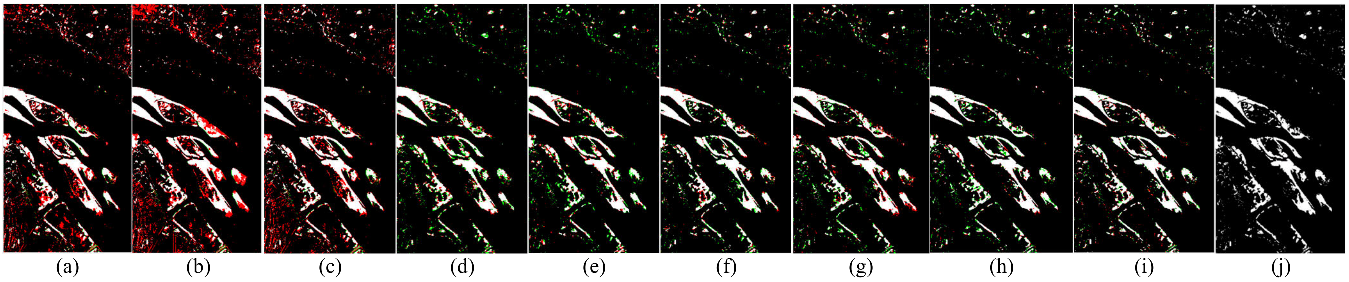


Fig. 3. Extracted change maps on the River Dataset from different methods of (a) AD, (b) CVA, (c) PCA-KM, (d) LSTM, (e) HybridSN, (f) 3-DCNN, (g) 2-DCNN, (h) CSANet, and (i) CBANet in comparison to (j) Ground-truth map, where the false alarms and missing pixels are labeled in red and green.

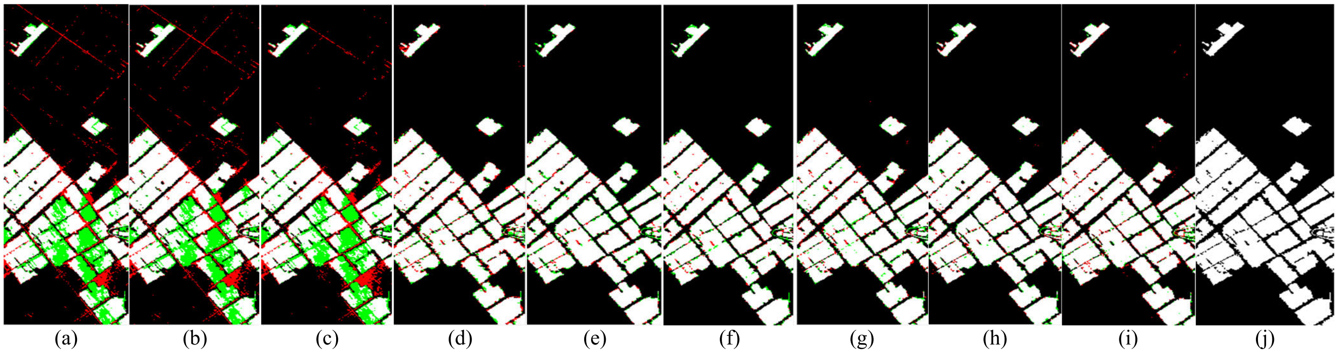


Fig. 4. Extracted change maps on the Yancheng Dataset from different methods of (a) AD, (b) CVA, (c) PCA-KM, (d) LSTM, (e) HybridSN, (f) 3-DCNN, (g) 2-DCNN, (h) CSANet, and (i) CBANet in comparison to (j) Ground-truth map, where the false alarms and missing pixels are labeled in red and green.

ones, as these are region-wise classification methods and more robust to spatial noise than pixel-wise unsupervised approaches. Although the OA from LSTM and HybridSN is relatively high, their CA is even lower than that of the unsupervised methods, leading to poor detection results in Fig. 4(d) and (e), especially the boundaries of the changing areas. For 2-D-CNN and 3-D-CNN, they have produced similar OA and KP as LSTM and HybridSN, though the visual results seem slightly better, although the connected changing region in the middle of the maps cannot be well distinguished. CSANet has yielded almost the same OA, AA, and KP as our proposed CBANet, where both of them are the top-performed models. However, our CBANet has a higher CA than the CSANet in the detection of the changed pixels, whilst the false alarms as indicated by NCA remain very comparable.

3) Experiments on the Hermiston Dataset: For the Hermiston dataset, the HCD results are shown and compared with the ground truth in Fig. 5 and Table IV, where the changing areas are mainly crop regions with simple round shapes. The results of the quantitative analysis are presented in Table IV. OA of all methods has achieved at least 97% or over 99% for DL methods. However, unsupervised methods have still produced quite a few false alarms, leading to lower UCA and KP values than all DL methods. For DL-based methods, as highlighted in Fig. 5, LSTM and HybridSN fail to accurately segment the edges of the changing area, where the boundaries of crop regions are connected together. Though 2-D-CNN and 3-D-CNN have slightly better results than LSTM and HybridSN, it is still difficult for them to detect crop regions with smooth edges. On the contrary, our CBANet can much more accurately detect the changing areas, with the KP

TABLE III
QUANTITATIVE ASSESSMENT OF DIFFERENT METHODS ON THE YANCHENG DATASET

	OA	CA	NCA	AA	KP
AD	0.8780	0.7494	0.9365	0.8429	0.7074
CVA	0.8755	0.7529	0.9312	0.8421	0.7025
PCA-KM	0.8828	0.7519	0.9424	0.8471	0.7180
LSTM	0.9555±0.0010	0.9246±0.0042	0.9702±0.0011	0.9472±0.0016	0.8967±0.0023
HybridSN	0.9641±0.0021	0.9350±0.0191	**0.9790**±0.0042	0.9555±0.0052	0.9160±0.0055
3-D-CNN	0.9665±0.0015	0.9427±0.0058	0.9774±0.0016	0.9601±0.0025	0.9221±0.0035
2-D-CNN	0.9667±0.0014	0.9413±0.0037	0.9781±0.0016	0.9603±0.0026	0.9223±0.0030
CSANet	**0.9715**±0.0009	0.9584±0.0015	0.9774±0.0020	0.9677±0.0003	**0.9335**±0.0023
CBANet	0.9713±0.0006	**0.9605**±0.0070	0.9768±0.0041	**0.9679**±0.0019	0.9332±0.0014

TABLE IV
QUANTITATIVE ASSESSMENT OF DIFFERENT METHODS ON HERMISTON DATASET

	OA	CA	NCA	AA	KP
AD	0.9728	0.9781	0.9367	0.9574	0.8824
CVA	0.9781	0.9843	0.9351	0.9597	0.9035
PCA-KM	0.9789	0.9858	0.9322	0.9590	0.9068
LSTM	0.9901±0.0010	0.9602±0.0074	0.9945±0.0009	0.9773±0.0036	0.9555±0.0046
HybridSN	0.9893±0.0006	0.9580±0.0014	0.9939±0.0011	0.9759±0.0047	0.9519±0.0030
3-D-CNN	0.9919±0.0003	0.9728±0.0081	0.9948±0.0014	0.9834±0.0033	0.9638±0.0016
2-D-CNN	0.9912±0.0004	0.9662±0.0077	0.9949±0.0012	0.9806±0.0033	0.9606±0.0021
CSANet	0.9923±0.0006	**0.9747**±0.0075	0.9950±0.0003	0.9848±0.0037	0.9659±0.0031
CBANet	**0.9928**±0.0010	0.9745±0.0057	**0.9955**±0.0007	**0.9850**±0.0024	**0.9678**±0.0008

0.40%–0.72% higher than that of 3-D-CNN and 2-D-CNN yet a much-reduced STD by 0.0008–0.0013, again validating the high efficacy of the proposed approach in HCD. In this dataset, although DL methods outperform all unsupervised ones with fewer false alarms and missing detection, the difference in terms of quantitative assessments is small, due mainly to the relatively simple background hence less noise caused by false alarms. Within the DL methods, LSTM has the poorest performance, whilst the results from 2-D-CNN and 3-D-CNN are quite similar. As the combination of 2-D-CNN and 3-D-CNN, HybridSN can extract spectral–spatial features of local regions, yet it fails to feature changing pixels. Also, it may suffer from overfitting due to too many convolutional layers contained. In addition, these three CNN models suffer from dealing with sporadic and isolated changing pixels because the large perceptual field in their convolutional layers can help to extract the global features but neglect small details. Again, CSANet and our CBANet have about the same results in terms of OA, AA, and KP, though it has a slightly higher CA than CBANet. In addition, it is worth noting that in all three datasets, the proposed CBANet has a (slightly) higher AA than the CSANet, indicating its strong capability in characterizing both large and small features for their accurate detection.

For our CBA model, however, the cross-band feature extraction module can extract the representative spectral feature whilst reducing the spectral dimension. The 2-D self-attention module can further fuse the spatial and spectral features for distinguishing both sporadic changing pixels and large changing areas. As a result, our proposed CBANet can consistently produce the best results than other benchmarking methods on all three datasets.

D. Ablation Experiments

1) Hyperparameter Analysis: To further validate the efficiency of our proposed CBANet, we compare the hyperparameters, the number of floating-point operations (FLOPs), and the overall running time in minutes (m) on the River dataset in Table V, including both the training time and testing time. It can be observed that HybridSN, 3-D-CNN, and 2-D-CNN, CSANet have much more hyperparameters, resulting in longer running time than our proposed method. Although LSTM has less running time and fewer hyperparameters, it has the worst detection accuracy of the three datasets. Thanks to the 1×1 convolutional kernel in the cross-band feature extraction module and 2-D self-attention module, our proposed CBA model can be considered a lightweight model which has fewer hyperparameters but performs better than other benchmarking methods.

2) Training Ratio Analysis: To fully validate the effectiveness of our proposed model, we evaluate the results of all the above-mentioned DL-based methods on the River dataset when the training pixels vary from 10% to 70%. As seen in Fig. 6, more training pixels will make the DL methods achieve better detection accuracy. Meanwhile, our CBANet can consistently achieve the highest OA and KP, where the

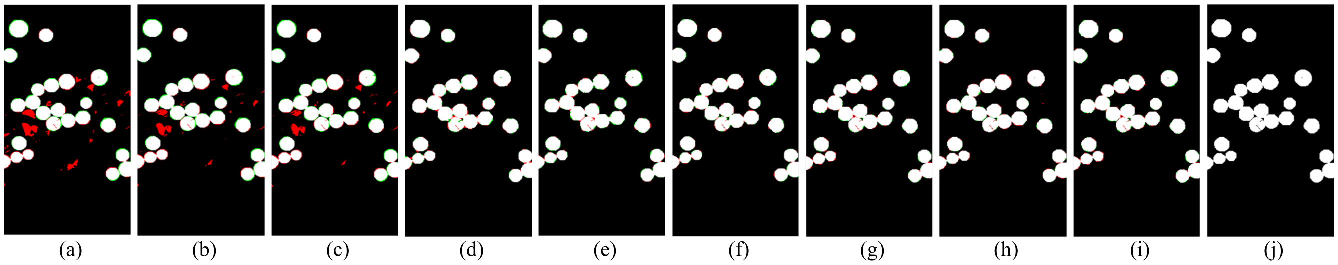


Fig. 5. Extracted change maps on the Hermiston Dataset from different methods of (a) AD, (b) CVA, (c) PCA-KM, (d) LSTM, (e) HybridSN, (f) 3-DCNN, (g) 2-DCNN, (h) CSANet, and (i) CBANet in comparison to (j) Ground-truth map, where the false alarms and missing pixels are labeled in red and green.

TABLE V
COMPARING THE PARAMETERS OF DIFFERENT DL-BASED METHODS ON RIVER DATASET

	LSTM	HybridSN	3-D-CNN	2-D-CNN	CSANet	CBANet
Hyperparameters (k)	213.79	5128.74	1613.03	607.43	2452.88	319.36
FLOPs (M)	3.51	1597.24	215.35	368.21	144.44	6.66
Running Time (m)	10.21	130.22	55.90	32.42	101.43	20.71

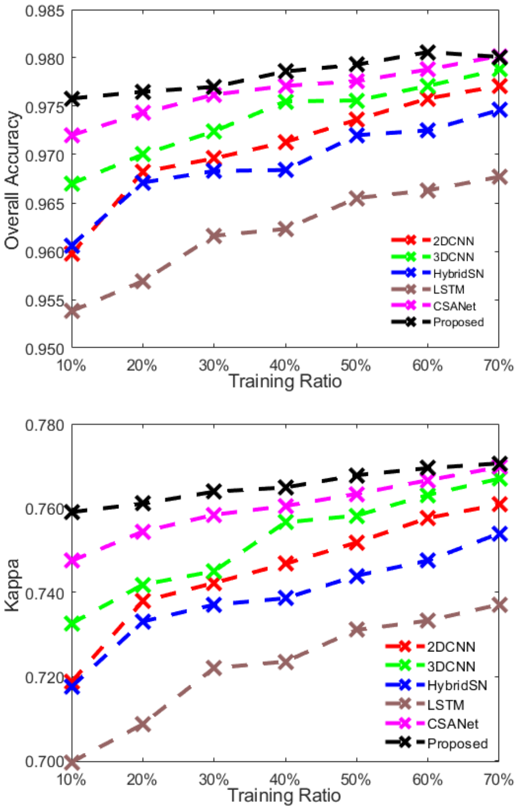


Fig. 6. Comparing the (top) OA and (bottom) KP results of all DL methods on the River dataset under different training ratios.

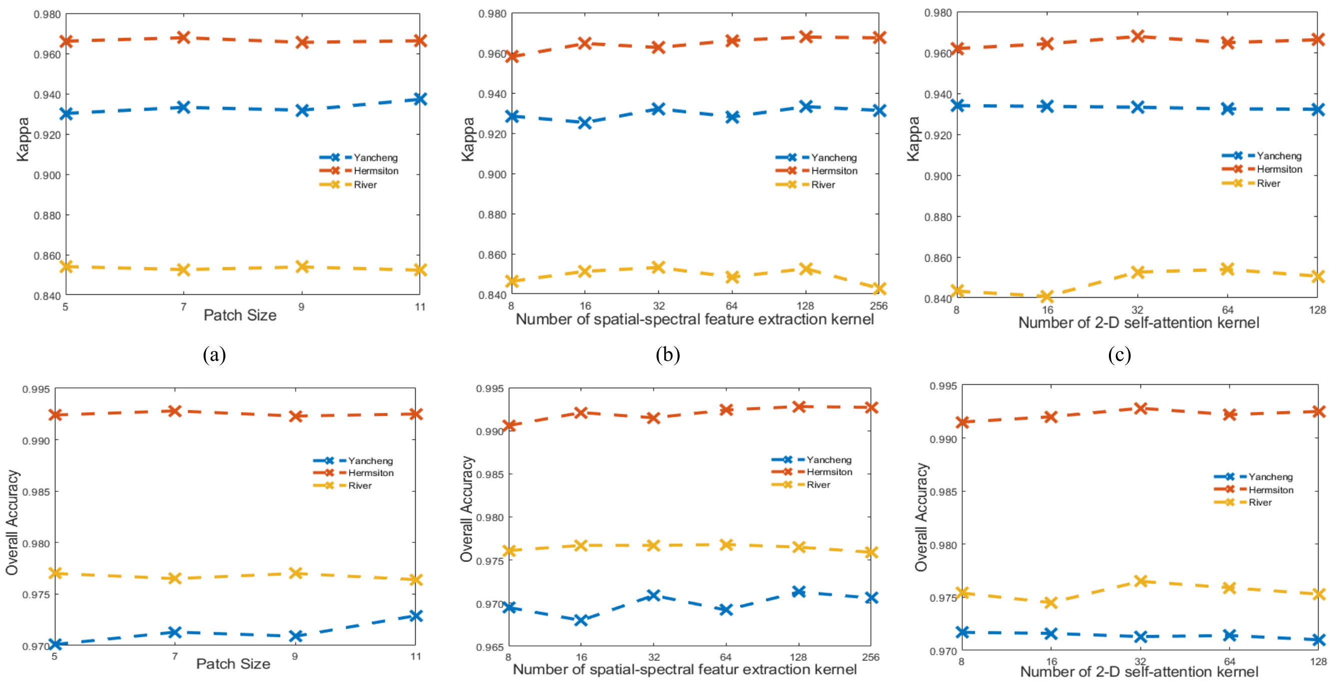
best OA and KP on the River dataset can reach 98.01% and 0.8765%, respectively.

3) Patch Size: We tested four patch sizes of {5 × 5, 7 × 7, 9 × 9, and 11 × 11} to analyze their impact on the CBANet. As shown in Fig. 7(a) and (d), an increase in patch size has a very limited effect on the KP and OA achieved when other module parameter settings are unchanged, though the largest patch size of 11 × 11 seems to have slightly improved OA and KP value on Yancheng dataset. That is why we choose the patch size of 7 × 7 in our experiments for all three datasets to balance between the detection accuracy and computational efficiency.

4) Number of spatial–spectral Feature Extraction Kernels: To find the optimal number of kernels in the spatial–spectral module, six different settings of {8, 16, 32, 64, 128, 256} are tested. As shown in Fig. 7(b) and (e), all three datasets have the highest OA and KP value when the number of kernels is 128. Therefore, we set the kernel number for spatial–spectral feature extraction as 128 throughout this article.

5) Number of 2-D Self-Attention Kernels: We have also evaluated the selection of the number of 2-D self-attention kernels, where the experiments on five different settings of {8, 16, 32, 64, 128} are conducted. As shown in Fig. 7(c) and (f), the variation trends of OA and KP value on the three datasets increase first and then decrease with the increasing number of kernels, and the classification result has the highest OA and KP value when kernel number of the 2-D self-attention module is 32.

6) Key Stage Analysis: In this section, compared with the traditional self-attention mechanism with 1 × 1 kernel, the efficacy of the 1-D convolution module with 1 × 3 or 3 × 1 kernel, 2-D self-attention module with 5 × 5 kernel, and the proposed 2-D self-attention module with 3 × 3 kernel is compared. As seen in Table VI, the 3 × 3 kernel outperforms other 1-D and 2-D kernels in the self-attention module. Specifically, for the River dataset, the OA and KP from the 3 × 3 kernel are 0.15% and 1% higher than those from the 1 × 1 kernel, respectively. Meanwhile, compared with the 1 × 1 kernel, the standard deviation of the OA and KP in the 3 × 3 kernel has been reduced by 53.8% and 56.6%, respectively. For the Yancheng and Hermiston datasets, an interesting finding is that the 5 × 5 kernel produces the worst results than all others. The possible reason is that this kernel is too large for the connected changed regions contained in these two datasets. In addition, a larger kernel in the 2-D



(a) (b) (c)

(d) (e) (f)

Fig. 7. Ablation experiments and results of the CBANet in different settings on the three datasets, including (a) Kappa values of different patch sizes, (b) different number of spatial–spectral feature extraction kernel, (c) different number of 2-D self-attention kernel, and (d) OA values under different patch size, (e) different number of spatial–spectral feature extraction kernel, and (f) different number of 2-D self-attention kernel.

TABLE VI
Comparison of OA and KP From Three Datasets With Various Kernel Sizes

	Kernel size	1×1	1×3	3×1	3×3	5×5
River	OA	0.9750±0.0026	0.9759±0.0029	0.9756±0.0009	**0.9765**±0.0012	0.9757±0.0008
	KP	0.8426±0.0083	0.8467±0.0005	0.8471±0.0031	**0.8526**±0.0036	0.8452±0.0051
Yancheng	OA	0.9707±0.0005	0.9700±0.0009	0.9711±0.0007	**0.9713**±0.0006	0.9689±0.0013
	KP	0.9319±0.0011	0.9301±0.0024	0.9325±0.0018	**0.9332**±0.0014	0.9276±0.0029
Hermiston	OA	0.9916±0.0014	0.9914±0.0009	0.9923±0.0006	**0.9928**±0.0010	0.9910±0.0006
	KP	0.9628±0.0017	0.9616±0.0039	0.9656±0.0016	**0.9678**±0.0008	0.9596±0.0027

self-attention module will inevitably lead to higher FLOPS and significantly more hyperparameters. In summary, the 2-D self-attention module with the 3 × 3 kernel can provide more accurate and robust results than other kernel sizes we have compared for HCD.

V. Conclusion

In this article, a novel lightweight end-to-end DL-based network, namely CBANet, is proposed for HCD. With the CBANet, the proposed cross-band feature extraction module has shown very good performance to fully extract and fuse the spectral information from bi-temporal HSI data whilst using the 1 × 1 kernels in the convolutional layer for efficiency. In addition, the proposed 2-D self-attention module has helped to capture deep spatial–spectral features for improving the feature representation and discrimination capabilities. The experiments on three publicly available HCD datasets have shown that the proposed CBANet outperforms other benchmarking models and has better stability and lighter weight than benchmarking DL models. This has fully validated the effectiveness and efficiency of the proposed model for the HCD task.

There are still some limitations to our proposed method. For example, the NCA of CBANet is inferior to some DL methods on the River dataset. To further improve the NCA, the detection of edge pixels of changed areas would be the key. The potential solution to achieving this purpose would be to add skip connections and multi-scale feature extraction layers in the model. Meanwhile, other advanced techniques such as superpixel [49], [50] and U-Net [51], [52] can be also employed to further improve the current model, especially in HCD accuracy.

References

[1] E. M. Domínguez, E. Meier, D. Small, M. E. Schaepman, L. Bruzzone, and D. Henke, "A multisquint framework for change detection in high-resolution multitemporal SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3611–3623, Jun. 2018.

[2] Y. Yan, J. Ren, Q. Liu, H. Zhao, H. Sun, and J. Zabalza, "PCA-domain fused singular spectral analysis for fast and noise-robust spectral–spatial feature mining in hyperspectral classification," *IEEE Geosci. Remote Sens. Lett.*, early access, Oct. 19, 2021, doi: 10.1109/LGRS.2021.3121565.

[3] D. Landgrebe, “Hyperspectral image data analysis,” *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[4] Y. Yan et al., “Non-destructive testing of composite fiber materials with hyperspectral imaging—Evaluative studies in the EU H2020 FibreEUse project,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[5] M. Fauvel, Y. Tarabalka, J. Atli Benediktsson, J. Chanussot, and J. C. Tilton, “Advances in spectral–spatial classification of hyperspectral images,” *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[6] G. M. Gandhi, S. Parthiban, N. Thummalu, and A. Christy, “NDVI: Vegetation change detection using remote sensing and GIS—A case study of Vellore district,” *Proc. Comput. Sci.*, vol. 57, pp. 1199–1210, Jan. 2015.

[7] P. Washaya, T. Balz, and B. Mohamadi, “Coherence change-detection with Sentinel-1 for natural and anthropogenic disaster monitoring in urban areas,” *Remote Sens.*, vol. 10, no. 7, p. 1026, Jun. 2018.

[8] B. Pengra, A. Gallant, Z. Zhu, and D. Dahal, “Evaluation of the initial thematic output from a continuous change-detection algorithm for use in automated operational land-change mapping by the U.S. Geological survey,” *Remote Sens.*, vol. 8, no. 10, p. 811, Oct. 2016.

[9] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, “Change detection in synthetic aperture radar images based on deep neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[10] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, “Deep learning-based change detection in remote sensing images: A review,” *Remote Sens.*, vol. 14, no. 4, p. 871, Feb. 2022.

[11] W. Sun and Q. Du, “Hyperspectral band selection: A review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019.

[12] A. Villa, J. Chanussot, J. A. Benediktsson, C. Jutten, and R. Dambreville, “Unsupervised methods for the classification of hyperspectral images with low spatial resolution,” *Pattern Recognit.*, vol. 46, no. 6, pp. 1556–1568, Jun. 2013.

[13] G. W. Humphreys and J. Sui, “Attentional control and the self: The self-attention network (SAN),” *Cognit. Neurosci.*, vol. 7, nos. 1–4, pp. 5–17, Oct. 2016.

[14] P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia, “Fusion of difference images for change detection over urban areas,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1076–1086, Aug. 2012.

[15] W. A. Malila, “Change vector analysis: An approach for detecting forest changes with Landsat,” in *Proc. LARS Symposia*, 1980, p. 385.

[16] C. Yang, J. H. Everitt, and J. M. Bradford, “Yield estimation from hyperspectral imagery using spectral angle mapper (SAM),” *Trans. ASABE*, vol. 51, no. 2, pp. 729–737, 2008.

[17] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, “PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data,” *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.

[18] J. Wang and C.-I. Chang, “Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.

[19] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, “Classification of hyperspectral images with regularized linear discriminant analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

[20] A. A. Nielsen, K. Conradsen, and J. J. Simpson, “Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies,” *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998.

[21] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, Dec. 1936.

[22] 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.

[23] B. Zhang, W. Yang, L. Gao, and D. Chen, “Real-time target detection in hyperspectral images based on spatial–spectral information extraction,” *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–15, Dec. 2012.

[24] 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. 2019.

[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] J. Qu, S. Hou, W. Dong, Y. Li, and W. Xie, “A multilevel encoder–decoder attention network for change detection in hyperspectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518113.

[27] X. Ou, L. Liu, B. Tu, G. Zhang, and Z. Xu, “A CNN framework with slow-fast band selection and feature fusion grouping for hyperspectral image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5524716.

[28] X. Li, Z. Yuan, and Q. Wang, “Unsupervised deep noise modeling for hyperspectral image change detection,” *Remote Sens.*, vol. 11, no. 3, p. 258, Jan. 2019.

[29] L. Wang, L. Wang, Q. Wang, and P. M. Atkinson, “SSA-SiamNet: Spectral–spatial-wise attention-based Siamese network for hyperspectral image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5510018.

[30] R. Song, W. Ni, W. Cheng, and X. Wang, “CSANet: Cross-temporal interaction symmetric attention network for hyperspectral image change detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[31] Y. Yang, J. Qu, S. Xiao, W. Dong, Y. Li, and Q. Du, “A deep multiscale pyramid network enhanced with spatial–spectral residual attention for hyperspectral image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5525513.

[32] M. Hu, C. Wu, and L. Zhang, “HyperNet: Self-supervised hyperspectral spatial–spectral feature understanding network for hyperspectral change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5543017.

[33] Y. Wang et al., “Spectral–spatial–temporal transformers for hyperspectral image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536814.

[34] 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.

[35] A. F. Agarap, “Deep learning using rectified linear units (ReLU),” 2018, *arXiv:1803.08375*.

[36] Z. Tong and G. Tanaka, “Hybrid pooling for enhancement of generalization ability in deep convolutional neural networks,” *Neurocomputing*, vol. 333, pp. 76–85, Mar. 2019.

[37] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, “Deep multiscale Siamese network with parallel convolutional structure and self-attention for change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406512.

[38] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” 2018, *arXiv:1803.02155*.

[39] J. Xia, Y. Cui, W. Li, L. Wang, and C. Wang, “Lightweight self-attention residual network for hyperspectral classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[40] G. Xie, J. Ren, S. Marshall, H. Zhao, R. Li, and R. Chen, “Self-attention enhanced deep residual network for spatial image steganalysis,” *Digit. Signal Process.*, early access, May 10, 2023, doi: 10.1016/j.dsp.2023.104063.

[41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[42] P. Henry, G. Chander, B. Fougnie, C. Thomas, and X. Xiong, “Assessment of spectral band impact on intercalibration over desert sites using simulation based on EO-1 hyperion data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1297–1308, Mar. 2013.

[43] H. Fu, A. Zhang, G. Sun, B. Shao, J. Ren, and X. Jia, “Unsupervised 3D tensor subspace decomposition network for hyperspectral and multispectral image spatial–temporal-spectral fusion,” *IEEE Trans. Geosci. Remote Sens.*, early access, May 9, 2023, doi: 10.1109/TGRS.2023.3272669.

[44] S. T. Seydi, M. Hasanlou, and M. Amani, “A new end-to-end multi-dimensional CNN framework for land cover/land use change detection in multi-source remote sensing datasets,” *Remote Sens.*, vol. 12, no. 12, p. 2010, Jun. 2020.

[45] R. Chen et al., “Rapid detection of multi-QR codes based on multistage stepwise discrimination and a compressed MobileNet,” *IEEE Internet Things J.*, early access, Apr. 20, 2023, doi: 10.1109/JIOT.2023.3268636.

[46] N. He et al., “Feature extraction with multiscale covariance maps for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 755–769, Feb. 2019.

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

[48] Q. Liu, F. Zhou, R. Hang, and X. Yuan, “Bidirectional-convolutional LSTM based spectral–spatial feature learning for hyperspectral image classification,” *Remote Sens.*, vol. 9, no. 12, p. 1330, Dec. 2017.

[49] G. Sun et al., “SpaSSA: Superpixelwise adaptive SSA for unsupervised spatial–spectral feature extraction in hyperspectral image,” *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6158–6169, Jul. 2022.

[50] P. Ma et al., “Multiscale superpixelwise prophet model for noise-robust feature extraction in hyperspectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508912.

[51] J. Ren et al., “Effective extraction of ventricles and myocardium objects from cardiac magnetic resonance images with a multi-task learning U-Net,” *Pattern Recognit. Lett.*, vol. 155, pp. 165–170, Mar. 2022.

[52] H. Zhao et al., “SC2Net: A novel segmentation-based classification network for detection of COVID-19 in chest X-ray images,” *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4032–4043, Aug. 2022.

Yinhe Li received the B.E. degree in electronic and electrical engineering from Northeast Electric Power University, Jilin, China, and the University of Strathclyde, Glasgow, U.K., in 2020. He is currently pursuing the Ph.D. degree with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K.

His research interests include hyperspectral imagery and deep learning.

Jinchang Ren (Senior Member, IEEE) received the Ph.D. degree in electronic imaging from the University of Bradford, Bradford, U.K., in 2009.

He is currently a Professor of computing science with Robert Gordon University, Aberdeen, U.K. His research interests include hyperspectral imaging, image processing, computer vision, big data analytics, and machine learning.

Yijun Yan (Member, IEEE) received the Ph.D. degree in electrical and electronic engineering from the University of Strathclyde, Glasgow, U.K., in 2018.

He is currently a Research Fellow with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K. His research interests include computer vision, hyperspectral imagery, pattern recognition, and machine learning.

Qiaoyuan Liu received the Ph.D. degree in analysis and planning of intelligent environment from Northeast Normal University, Changchun, China, in 2019.

She is currently an Assistant Professor with the Changchun Institute of Optics, Precision Mechanics and Physics, Chinese Academy of Sciences, Changchun. Her research interests include visual tracking and image analysis.

Ping Ma received the Ph.D. degree from the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, U.K., in 2022.

She is currently a Research Fellow with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K. Her research interests include multimodal remote sensing, hyperspectral imaging, and machine learning.

Andrei Petrovski received the Ph.D. degree in computer science (computational optimization) from Robert Gordon University (RGU), Aberdeen, U.K., in 1998.

He is currently an Associate Professor with RGU, with more than 20 years’ experience of working in academia, teaching and doing research in cybersecurity, computer and sensor networks, computational intelligence, and machine learning.

Haijiang Sun received the Ph.D. degree in electronic engineering from the Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun, China, in 2012.

He is currently a Professor with the Perception and Display Laboratory, CIOMP. His research interests include target recognition and tracking, and high-definition video enhancement and display.