SSViT-HCD: A Spatial–Spectral Convolutional Vision Transformer for Hyperspectral Change Detection

Ayesha Shafique, Seyd Teymoor Seydi, Tayeb Alipour-Fard, Guo Cao, and Di Yang

***Abstract*—In recent decades, the wide use of deep-learning-based methods has consistently improved the performance of remote sensing images and is widely used for hyperspectral change detection (HCD) tasks. However, most of the existing HCD method is based on the convolutional neural network (CNN), which shows limitations in long-range dependencies and also cannot mine sequence features well. The change detection (CD) performance still has margins for improvement. In this study, inspired by the excellent performance of transformers in computer vision, which has shown a significant ability to model global dependencies to attenuate the loss of long-range information, we built a hybrid spatial–spectral convolutional vision transformer (SSViT) for HCD. Our proposed method combines the merits of CNN and transformer to fulfill effective and efficient HCD. This study focused on highly reliable pseudo sample data generation by a selection scenario. To generate a pseudo sample, we have used different methods: 1) predicting change and no-change areas by using Euclidean distance; 2) thresholding by the Chan–Vese segmentation method for determining change and no-change pixels for intensity maps; 3) sorting of change and no-change pixels; and 4) selection of the minimum value of initial no-change pixels as pseudo change sample data, in addition to choosing the maximum intensity value for change candidate pixels as change sample data. The highly reliable change pixels were selected, and then, pseudo training data were used to train the SSViT model. At last, the change map is generated by training the SSViT network based on pseudo training data. The performance of the SSViT model is evaluated for real-world hyperspectral (HS) datasets with different change land cover types. Furthermore, a new series of HS images is introduced for CD purposes. The results of CD show that the HS images have a high potential for detecting subtle changes. The experimental results demonstrate that the proposed SSViT could outperform the advanced HCD methods.**

***Index Terms*—Deep learning (DL), change detection (CD), hyperspectral image (HSI), vision transformer (ViT).**

I. Introduction

THE advancement of optical sensor technology over the past few decades has provided a great amount of information in terms of achieving required spatial, spectral, and temporal resolutions. The rich spectrum information included in hyperspectral images (HSIs), in particular, develops new application domains and raises new technological problems in remote sensing (RS) [1], [2], [3]. In particular, HSI is a 3-D data cube with 2-D spatial and one-dimensional spectral information.

Compared to multispectral images, HSI has high-resolution spectral and spatial information with hundreds of bands, optimizing the quality of HRS images [4]. Since HSIs include a plethora of spectral information, they have been extensively used in various domains, including image classification [5], [6], anomaly detection [7], [8], change detection (CD) [9], [10], and so on.

In RS, hyperspectral change detection (HCD) is the process of recognizing discrepancies between images captured over the same location at different times [11], [12]. Up to now, many traditional methods have been proposed for HCD: 1) algebra-based methods; 2) transformation-based methods; and 3) classification-based methods. The most prevalent approach to change vector analysis (CVA) is generally an unsupervised method that uses spectral vector subtraction [13]. In addition, various modified CVA approaches have been proposed, such as sequential spectral CVA, which overcomes the problems associated with the original CVA. A hybrid vector [14] was created using a change vector and spectral angle, and an adaptive fusion approach was used to create a CD map. Another development made to CVA was called robust CVA, which provides information on the intensity and kind of change, and robustness to changes in viewing geometry or registration noise [15]. However, CVA-based approaches have certain drawbacks; as the number of bands increases, it becomes more challenging to detect the change kinds and determine an adequate threshold [16]. Some methods based on image transformation, such as principal component analysis (PCA), exploit the variance in the principal components of the combined multitemporal HSIs [17]. PCA is mainly dependent on image statistics and is particularly vulnerable to unbalanced data. An unsupervised method of multivariate alteration detection (MAD) [18] for analyzing multitemporal image CD is based on canonical correlation analysis. The MAD approach can efficiently eliminate correlation, but the noise has

Manuscript received 17 October 2022; revised 18 December 2022; accepted 21 February 2023. Date of publication 2 March 2023; date of current version 21 July 2023. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191284, in part by the National Natural Science Foundation of China under Grant 61801222, and in part by the National Center for Atmospheric Research through the Computational and Information Systems Laboratory Research Computing Project under Grant WYOM0126. *(Corresponding author: Guo Cao.)*

Ayesha Shafique and Guo Cao are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: ayeshashafique@njust.edu.cn; caoguo@njust.edu.cn).

Seyd Teymoor Seydi and Tayeb Alipour-Fard are with the School of Surveying and Geospatial Engineering, University of Tehran, Tehran 1417935840, Iran (e-mail: seydi.teymoor@ut.ac.ir; tayebalipour@ut.ac.ir).

Di Yang is with the University of Wyoming, Laramie, WY 82071 USA (e-mail: dyang1@uwyo.edu).

Digital Object Identifier 10.1109/JSTARS.2023.3251646

a substantial influence on the results, and the threshold must be manually set. The improved iteratively reweighted MAD (IR-MAD) [19] creates a no-change background to identify changes. In particular, slow feature analysis (SFA) may also extract features from time series [20], [21]. SFA can be utilized for CD by suppressing invariant pixels and emphasizing those that are changing [22]. These approaches depend greatly on empirically generated algorithms to extract discriminative features, which are typically difficult to obtain meaningful results on HS images. Transformation-based approaches are helpful in terms of reducing dimensionality and noise, as well as emphasizing the change or no-change features associated with certain changes. Although the changes affect a large part of the imagery, it will take a large amount of time to create the new components. Furthermore, some methods are based on image classification, including post and direct classification. The first method classifies two images independently and then considers pixels or regions belonging to various classes as changes. For direct classification, multitemporal images are stacked together, using the same classifier to identify changed categories [23].

Even though the conventional methods demonstrated that the proposed methods are effective, several obstacles still need to be overcome. One of these obstacles is the fact that during the process of image acquisition, HSIs are easily influenced by noise brought on by the atmosphere. Over this, the low-level representations of HSIs are insufficiently discriminative for CD, making it crucial to reliably identify change regions for classical approaches.

The image quality directly affects the CD accuracy of HSI. Furthermore, HSI-CD approaches must deal with high dimensionality, computational load, and limited datasets. It is difficult to detect changes in high-dimensional feature spaces with hundreds of narrow continuous bands, so high dimensionality can increase the implicit and make it more difficult to distinguish changes [24]. To tackle these challenges, band-selection and feature-extraction approaches have been proposed; nevertheless, crucial spectral information may be lost when using these methods. Another major issue is the dataset limitation, as HSIs lack label information due to the difficulty of obtaining change information on real-world objects.

Fortunately, the advancement of deep learning (DL) offers a viable method of addressing these challenges. DL has steadily replaced various classical algorithms, acquiring an overwhelming superiority in CD. DL has been recognized to be a viable approach for dealing with high dimensionality. Similarly, DL can process complex HSIs data by efficiently extracting semantic features of images by reducing the dimensions of the data. The implementation of the convolutional neural network (CNN) facilitates the use of spectral data of pixel points and their adjacent pixels in the CD process and significantly overcomes the restriction of only using spectral information equivalent to a single-pixel point. In this regard, researchers extensively deployed the DL-based HSI-CD methods and obtained significant results. Yuan et al. [25] focused on the semisupervised CD technique and presented a distance metric learning technique for CD in a "noisy" state. Consequently, the presented technique outperforms in both "ideal" and "noisy" states for hyperspectral datasets. In addition, Liu et al. [26] proposed a CD approach based on spectral unmixing. Ertürk et al. [27] used dictionary pruning for sparse unmixing-based CD for HSI. The proposed method alleviates the unmixing process's ill-conditioning while also reducing computation time and improving CD performance. Wu et al. [28] introduced a joint sparse representation method for hyperspectral anomalous CD. These methods that are based on unmixing-based [26], [27] sparse representation [28] have obtained significant detection accuracy in HSI-CD. Despite the fact that these hyperspectral CD algorithms are widely used, issues remain, such as a lack of theoretical foundation and appropriate evaluation standards, a lack of detection method versatility, a lack of multisource data integration analysis, and low utilization of spatial information during detection.

In addition, the deep noise modeling method was also proposed by Li et al. [29] for CD in hyperspectral imagery datasets. To train discriminative features from the high-dimensional dataset, a fully convolutional network (FCN) was used; for fusing feature maps, a two-stream feature technique was deployed, and a noise modeling framework was utilized to deal with noise conditions. Besides, Wang et al. [24] used 2-D CNN-based techniques for HSI-CD in an end-to-end manner, where a new mixed affinity matrix is created and pixel change types are derived using CNN output. Particularly, Song et al. [30] presented a novel HSI-CD framework Re3FCN that employs an FCN with 3-D convolutional layers and a convolutional long short-term memory. In the first phase, PCA is used to select the training sample, and the second step is the use of recurrent CNN for training and testing. Their proposed method achieved a superior result to CD but relied significantly on training datasets. Seydi et al. [12] investigated a DL-based framework that relies on the image differencing method and the 3-D CNN to detect change areas and make a decision on detected areas, respectively. Moustafa et al. [31] presented a semantic-segmentation-based model and used ROS in preprocessing, DL, and bagging ensemble to manage unbalanced datasets. Their proposed method used four different types of Unet models to separate change and no-change zones.

Huang et al. [32] used a tensor-based DL method for HCD. However, the proposed method has drawbacks, such as mixed pixel problems and time complexity issues. Qu et al. [33] proposed the DL framework with a multilevel encoder–decoder attention network for HCD. The hierarchical features are fully utilized in the ML-EDAN framework for CD in HSIs. Furthermore, ML-EDAN was trained as an end-to-end framework and explored the reconstruction and pixelwise classification error. Qu et al. [34] developed the novel dual-branch change amplification framework based on the graph convolutional network for HSI-CD. The proposed network completely extracts and efficiently amplifies the change features of multitemporal HSI-CD.

In the literature, most of the HCD algorithms are based on CNN models. The CNN models are known as robust feature representation frameworks in RS. These models provided promising results in the many applications of RS. Although the CNN can capture local features (inductive bias) well, its inherent network structure does not allow it to mine and represent sequence attributes of spectral features. Deep semantic features cannot be acquired by a shallow convolution layer in the CNN frameworks. Instead, they capture deep semantic information with an increase in the depth of the model by adding convolution layers and feature map reduction size by pooling layers. Due to this fact, the CNN models suffer computational complexity. Transformers have proven useful for solving several vision problems in recent years due to their ability to capture long-range relationships and sequence-based image modeling. Although the transformer captures spectral signatures well, it is not powerful enough for capturing local semantic features or making use of spatial information. In contrast, the transformer's self-attention (SA) mechanism is effective at modeling global interactions between token embeddings, but local mechanisms for information exchange within local regions are lacking. To tackle the abovementioned limitation of the CNN model and transformer, we introduced a novel hybrid framework for HCD. We propose a spectral–spatial convolutional vision transformer (SSViT) model for HCD as a way of taking advantage of the transformer's ability and convolution layers to acquire local spatial semantic information and model the relationship between adjacent sequences.

To summarize, the most significant contributions are as follows.

1) We propose a hybrid SSViT-based method for HCD. Our proposed method has high efficiency and can provide reliable sample data and improve the CD result. According to our literature review, this is the first work to attempt to implement a hybrid convolutional vision transformer (ViT) for HCD.
2) For the first time, we used Chan–Vese segmentation for the HCD and a new Agriculture PRISMA dataset for the CD task.
3) We redesign the convolutional block attention module (CBAM) for HCD based on the 3-D structure of convolution and pooling layers.
4) Furthermore, several ablation studies have shown that the proposed framework is effective and capable of detecting subtle changes.
5) Our proposed model is compared with advanced DL-based methods such as GetNet, TDRD, PTCD, and ViT. The experimental results show that the suggested SSViT method outperforms the advanced CNN algorithm.

The rest of this article is organized as follows. Section II explains the methodology of our proposed SSViT model. Section III describes the datasets. Section IV includes extensive ablation studies as well as experimental comparisons of the SSViT model with several CNN models. Finally, Section V concludes this article.

II. Methodology

The implementation detail of the HCD method is shown in Fig. 1. As can be observed, the HCD consists of three primary steps: 1) data preparation and preprocessing; 2) pseudo sample generation; and 3) SSViT model training and tuning model parameters.

A. Data Preparation and Preprocessing

Hyperspectral imagery needs to be preprocessed before CD can be considered. The main preprocessing included no-data removal, smile correction, radiometric correction, and atmospheric correction. The mentioned preprocessing is a spectral correction related to the image's pixel value. The second category is a geometric correction, which refers to the spatial location of pixels in the bitemporal dataset.

B. Pseudo Sample Generation

In the proposed model, generating reliable samples is considered one of the essential aspects of HCD. To this end, many researchers have worked only on pseudo sample data generation without refining. Fig. 2 shows a schematic of the histogram, a predictor for CD that there is an uncertain part in the result of thresholding. This uncertainty originated from mixing change and no-change pixels due to some conditions (i.e., noise effect, atmospheric conditions, and complexity of objects). Since the segmentation algorithm cannot discriminate the change pixels from the background, refining sample data is vital. To this end, this research introduces a simple and effective solution with a low computational cost. Thus, this study focused on highly reliable pseudo sample data generation by a selection scenario. The increasing reliability of sample data can improve the results of the HCD. The pseudo sample generation consists of four parts: 1) prediction of change and no-change areas by the Euclidean distance (EU); 2) binary segmentation by the Chan–Vese method for determining change and no-change pixels for intensity map (EU); 3) sorting of change and no-change pixels based on their magnitude in EU prediction results; and 4) selection of the minimum value of initial no-change pixels as pseudo change sample data, addition to choosing maximum intensity value for change candidate pixels as change sample data. The number of pixels is a direct dependence on the dataset and efficiency of the predictor, which can be determined based on knowledge.

1) EU Algorithm: There are numerous algorithms available for predicting change and no-change areas [35]. The EU is one of the most common metrics for predicting change and no-change regions in bitemporal datasets. Unlike other CD predictors, such as PCA, MAD, and IR-MAD, in which high-order statistical structures are used, the EU is simple to implement and has a low degree of complexity compared to other CD predictors. In addition, the EU algorithm can better discriminate between change (foreground) and no-change areas (background) in comparison with other algorithms such as the spectral angle mapper. Furthermore, this predictor uses the L2 norm, which can be used to discriminate change areas from the background (no-change area) as well. The intensity value of changes for two pixels of bitemporal HSI can be defined as follows:

$$\mathrm{EU}^{C} = \left(X_{i,j}^{c} - Y_{i,j}^{c}\right)^{2} \tag{1}$$

$$\text{Intensity} = \sum_{c=1}^{N} \left(X_{i,j}^{c} - Y_{i,j}^{c}\right)^{2} \tag{2}$$

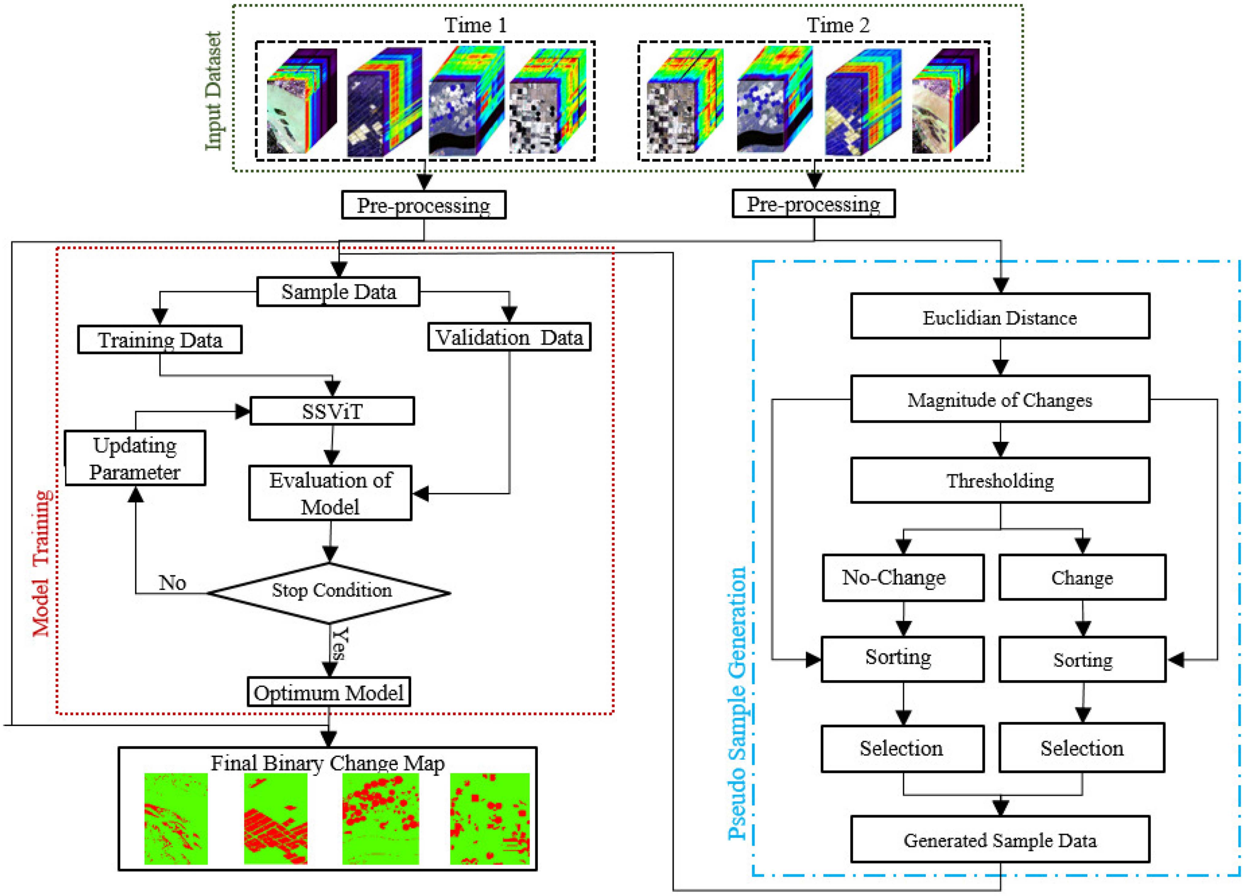


Fig. 1. General framework of HCD.

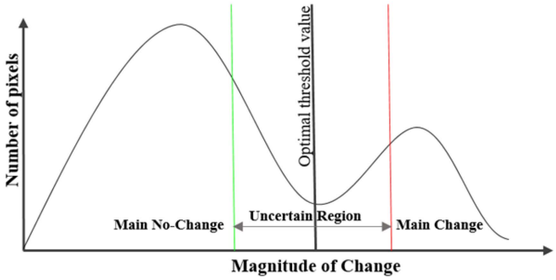


Fig. 2. Schematic of the change threshold.

where $X_{i,j}^{c}$ and $Y_{i,j}^{c}$ are the first and second time of bitemporal HSI in row i and column j in the band c, respectively. Furthermore, N shows the number of spectral bands in the bitemporal HSI.

2) Chan–Vese Segmentation: The edge-based segmentation methods are used widely for many applications in image processing that use image gradient information. However, these methods have low complexity, but objects with weak and blurred boundaries cannot be detected using these methods. The Chan–Vese method [36] is a typical and popular region-based method based on Mumford–Shah segmentation that is a flexible and powerful model for active contours in the segmenting of types of images [37]. This model is based on an energy minimization problem and ignores edges completely (image gradient); instead, it fits a two-phase piecewise-constant model optimally to the input data [38]. Because the method assumes that image intensity values are constant within each foreground and background region, it can effectively segment images with homogeneous intensity values across areas [39].

3) Sorting of Candidate Pixels: The candidate change and no-change pixels are sorted based on intensity value. This process consists of two primary steps: 1) applying a mask of the change and no-change on the intensity map and 2) extracting intensity values for change and no-change areas and arranging them based on intensity values.

4) Selection of Candidate Pixels: Due to some factors, such as atmospheric conditions and the existence of noise among bitemporal HSIs, there is a high mixing of change and no-change pixels. Thus, the classification of change and no-change pixels is difficult. The use of candidate change and no-change pixels without refining them leads to an effect on change results in supervised learning methods. To this end, we refine change and no-change pixels in a selection manner to increase their reliability. It is clear that the change pixels have a high value in the intensity map, while the no-change pixels contain the lowest intensity value in the initial change map. Thus, the highly reliable change pixels were selected for the next analysis.

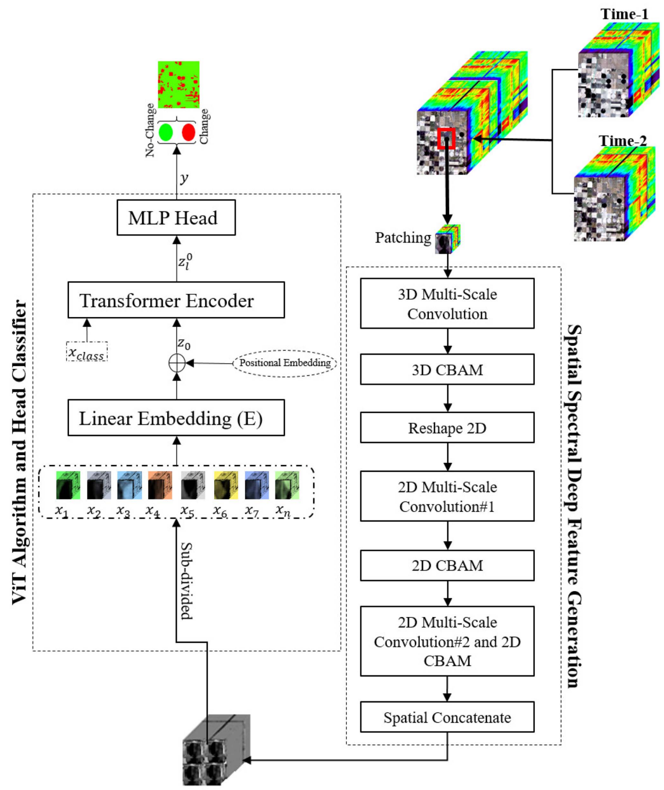

Time-1
Time-2
No-Change
Change
y
MLP Head
z_l^0
Transformer Encoder
x_{class}
z_0
Positional Embedding
Linear Embedding (E)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_n
Sub-divided
ViT Algorithm and Head Classifier
Patching
3D Multi-Scale Convolution
3D CBAM
Reshape 2D
2D Multi-Scale Convolution#1
2D CBAM
2D Multi-Scale Convolution#2 and 2D CBAM
Spatial Concatenate
Spatial Spectral Deep Feature Generation

Fig. 3. Main structure of SSViT for HCD.

C. SSViT Algorithm

The main structure of the proposed SSViT for HCD is illustrated in Fig. 3. As seen, SSViT has been built in two main parts: 1) shallow deep feature generation by hybrid 3-D/2-D convolution layers and attention mechanisms and 2) ViT for deep semantic feature generation and classification by classifier head.

First, the patch dataset is represented by a 3-D multiscale convolution block. Then, a 3-D CBAM attention module is utilized for informative feature generation. The extracted features reshape and change into the 2-D structure to be an explorer by 2-D convolution layers. Next, the 2-D multiscale convolution block (#1) is utilized to represent deep features also, and the 2-D CBAM mechanism is employed for more consideration. In the last step, the deep features are fed into the second 2-D multiscale convolution block (#2). The output of each 2-D convolution layer in this block is concatenated into a spatial dimensional. The output of the concatenation layer is transferred to the ViT algorithm for extracting deep semantic features. Then, the classifier head is utilized to make a decision.

D. Multiscale Convolution Block

CNN-based frameworks rely heavily on convolution layers to generate deep meaningful features. The size of kernel convolution has a key role in the extraction of deep features in that multiscale convolution block, which utilizes the convolution layers with various kernel sizes. The main advantage of the multiscale convolution block is that it enhances the robustness of the SSViT against variations in the size change objects. This research combines the 3-D and 2-D convolution layers. Hyperspectral imagery has high content spectral information that utilizes 3-D convolution to help capture spectral information among spectral bands. After convolution layers, the CBAM attention mechanism block is utilized to generate more informative features. The main structure of the 3-D/2-D multiscale convolution block is shown in Fig. 4.

E. CBAM Attention Module

The main task of the attention mechanism is to emphasize deep meaningful features. The CBAM module is a popular and effective attention module that can learn “what” and “where” to attend to the channel and spatial dimensions of feature maps, respectively. The CBAM module has two branches that are included: 1) spatial attention module and 2) channel attention module. The main structure of the CBAM module is shown Fig. 5. The channel attention module tries to extract the channels that include the informative spectral information. To this end, two global average/max-pooling layers are employed [see Fig. 5(a)]. Then, the two multilayer perceptron (MLP) layers with a reduction rate are used to minimize the computational cost and model parameters. Finally, the obtained features are fed to the Sigmoid activation function to generate the channel attention map.

The spatial attention module emphasizes parts containing key information. In this regard, the average/max-pooling layers are applied. Afterward, a convolution layer with Sigmoid activation functions generates the spatial attention map, as shown in Fig. 5(b). The 3-D CBAM module follows the abovementioned process, but it uses the 3-D pooling and convolution layers. Furthermore, the 2-D CBAM module has the same process while using the 2-D structure of pooling and convolution layers.

F. ViT Algorithm

Transformers have proven useful for solving several vision problems in recent years due to their ability to capture long-range relationships and sequence-based image modeling. ViT is the most popular and standard transformer-based method that can provide promising results in many applications of RS. This method uses the transformer model instead of convolution layers. This algorithm consists of three main components: 1) a patch embedding layer; 2) a transformer encoder; and 3) a head classifier.

In the first step, the generated deep features by the first part are divided into nonoverlapping patches based on the defined patch size. These patches are fed into the transformer encoder part and viewed by the transformer as individual tokens. The positional encoding is incorporated, which forces the model to use the order of the sequence and position information. To generate positional vectors, sine and cosine functions of different frequencies are incorporated. The generated patches are mapped into a feature vector of the model dimension d using a learned embedding matrix.

For the classification task, the embedded vector must be concatenated with a learnable classification token layer. As the last step, a trainable embedding tensor ($E_{\text{positional}}$) is added to

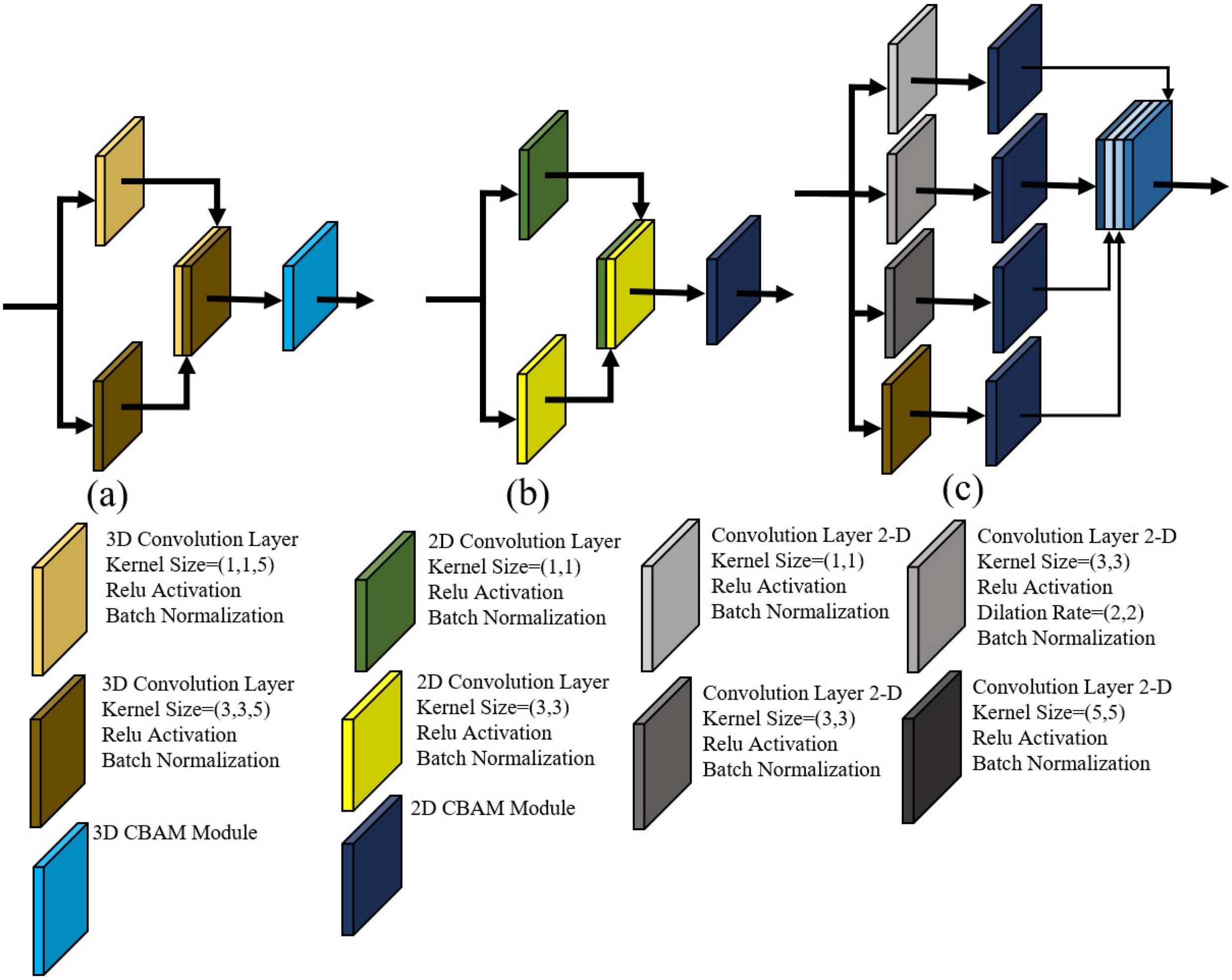


Fig. 4. Multiscale convolution blocks. (a) 3-D multiscale convolution block. (b) 2-D multiscale convolution block (#1). (c) 2-D multiscale convolution block (#2).

the concatenated projection sequence. The output of the patch embedding layer with the token z_0 is given as

$$z_o = [x_{\text{class}}; x_1 E; x_2 E; \ldots; x_n E] + E_{\text{positional}} \tag{3}$$

where x_i is a linearly projected feature map. The output of the patch embedding layer is used as the input of L transformer encoder layers for feature extraction. The transformer encoder layer extracts more abstract features from the embedded patches. The transformer layer has two main components that include a multihead self-attention layer (MSA) and an MLP. Furthermore, there are subcomponents in the transformer block that include: 1) the normalization layer used to stabilize hidden state dynamics and 2) the residual connections employed to prevent the vanishing gradient problem. A central component of the encoder transformer part in Fig. 6 is the MSA block, which includes many SA layers running in parallel. The purpose of the SA layer is to capture the interaction among all the embedding entities by encoding each entity. Consequently, each entity is the weighted sum of all entities in the sequence, where the weights are based on the attention scores. The MSA mechanism in every layer includes SA blocks to enable the encapsulation of multiple complex relationships between different elements in the sequence. The result of SA blocks in the MSA layer concatenates into a single matrix. The output of the lth MSA layer can be calculated as

$$z'_l = \text{MSA}\left(\text{LN}(z_{l-1})\right) + z_{l-1}, \quad l = 1, \ldots, L \tag{4}$$

where LN() is the normalization layer and z_l is the encoded image representation. Next, each encoder block is followed by a fully connected feedforward dense block that is estimated as

$$z_l = \text{MLP}\left(\text{LN}\left(z'_l\right)\right) + z'_l, \quad l = 1, \ldots, L. \tag{5}$$

In the final layer of the encoder, we take the first element in the sequence and pass it to a classifier to predict the class label. As seen, Fig. 7 illustrates the main structure of the classifier head

$$= \text{MLP}\left(z_l^0\right) \tag{6}$$

where the MLP component has two fully connected layers with the sigmoid activation function.

G. Training Process

The training process is applied in an iterative manner by estimating the error of the network by the loss function. For the training model, the generated pseudo sample dataset is divided into three parts are included: 1) training dataset (65%);

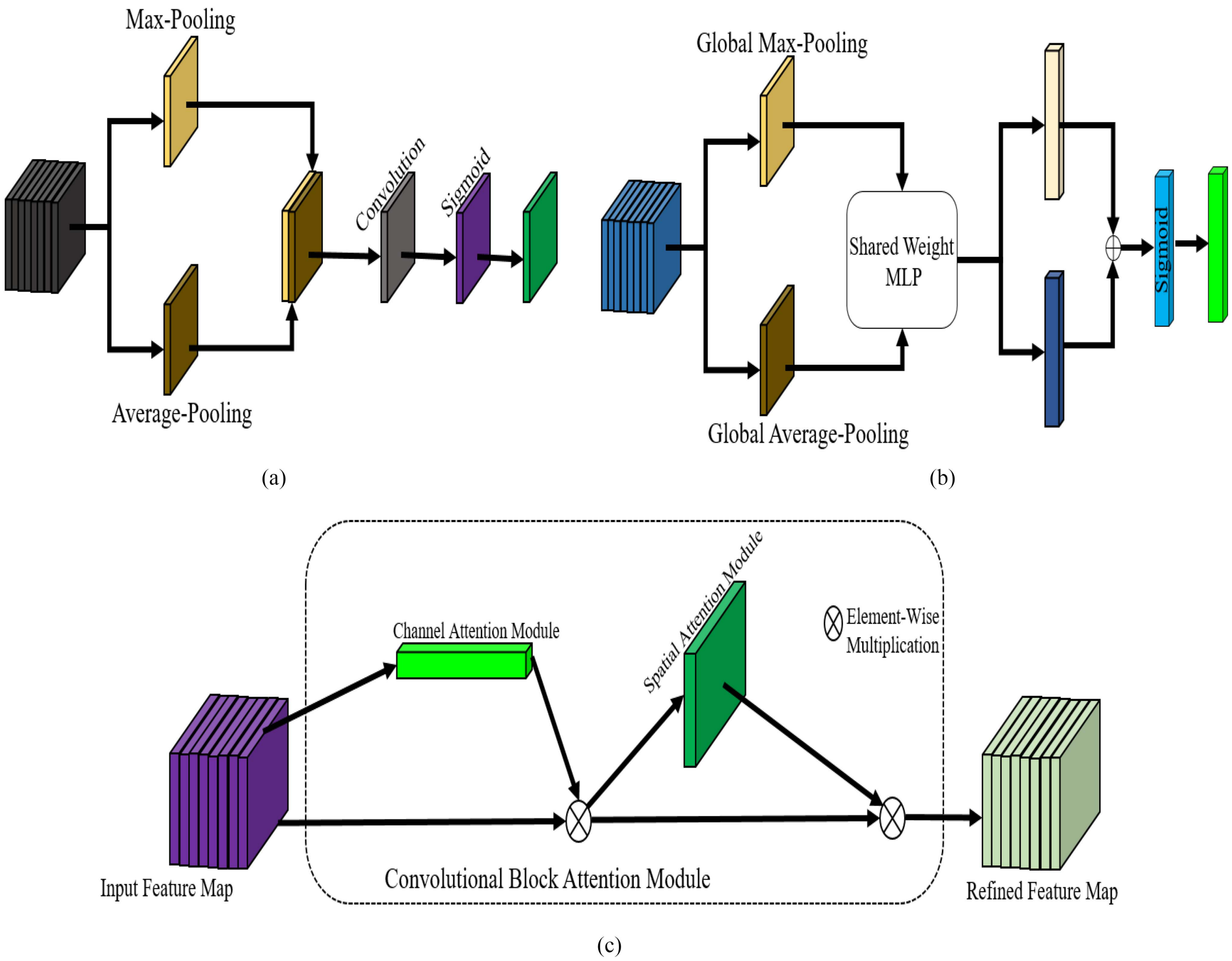


Fig. 5. General overview of the CBAM attention module. (a) CBAM module, (b) Channel attention module. (c) Spatial attention module.

2) validation dataset (15%); and 3) testing dataset (20%). The model is trained by a training dataset, while the error of the network is evaluated on the validation dataset by the loss function. Finally, the model is assessed on the testing dataset by quantity measurement indices such as overall accuracy (OA). Due to backpropagation, the model parameters are tuned by employing an optimizer that the adaptive moment estimation optimizer has employed. The cost function was also cross entropy

$$H_p(q) = \frac{-1}{N}\sum_{i=1}^{N} y_i \log\left(p(y_i)\right) + (1 - y_i)\,.\log\left(1 - p(y_i)\right) \tag{7}$$

where y is a label, $p(y)$ is the predicted probability observation, and N is the number of classes.

III. Datasets

The hyperspectral CD evaluation was carried out utilizing the following different hyperspectral datasets obtained by the Hyperion sensor.

1) *River dataset*: The dataset images were captured on May 3, 2013 and December 31, 2013 in Jiangsu Province, China. The size of the pixel is 436 × 241. The dataset images are shown in Fig. 8(a) and (b).
2) *Hermiston dataset*: This dataset was captured on May 1, 2004 and May 8, 2007 near Hermiston City in Umatilla County, OR, USA. The size of the pixel is 306 × 241. The Hermiston dataset images are shown Fig. 9(a) and (b).
3) *Farmland dataset*: This dataset acquired farmland around Yuncheng, Jiangsu Province, China. The data were collected on May 3, 2006 and April 23, 2007. The images have a spatial resolution of 30 m and a spectral resolution of 10 nm. Fig. 10(a) and (b) represents the dataset images.
4) *Agriculture PRISMA dataset:* PRISMA is an EO system with unique electrooptical equipment that combines a hyperspectral sensor with a medium-resolution panchromatic camera and is completely supported by the Italian Space Agency. The PRISMA sensor captured continuous hyperspectral datasets with a 29-day repeated orbital period. In this study, we choose the dataset size of 312 × 349 pixels, spectral band 169, and spatial size 30 m. Fig. 11(a) and (b) represents the agriculture PRISMA dataset images.

IV. Results

This section compares the results of implementing our proposed method on four datasets with the ground truth (GT)

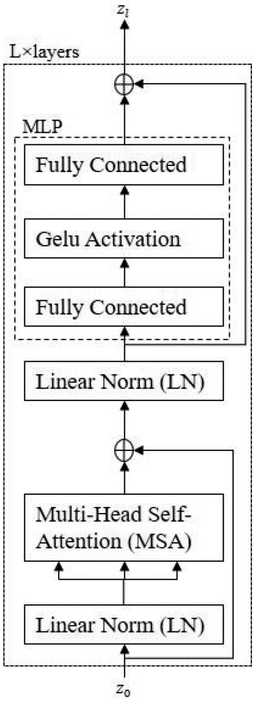


Fig. 6. Main transformer encoder.

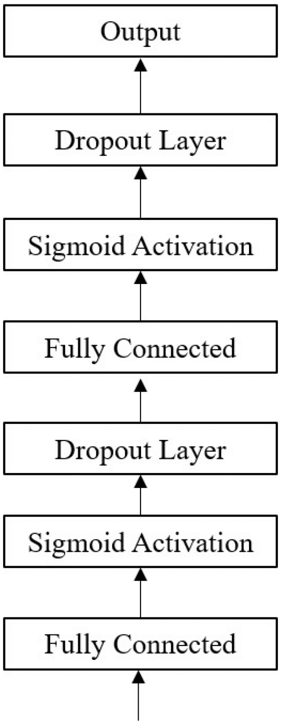


Fig. 7. MLP head classifier for binary CD.

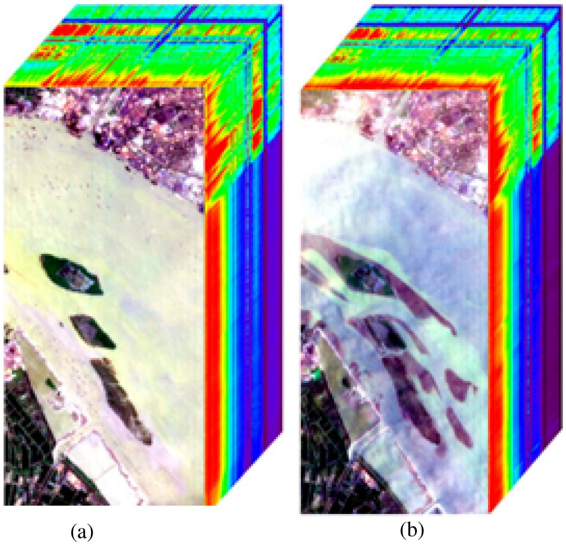


Fig. 8. River dataset. (a) River dataset image T1. (b) River dataset image T2.

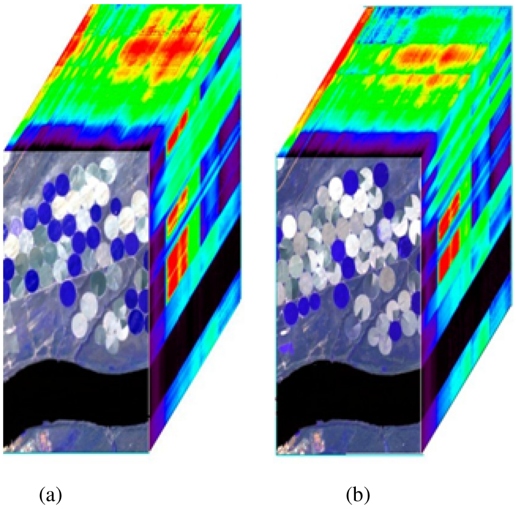


Fig. 9. Hermiston dataset. (a) Hermiston dataset image T1. (b) Hermiston dataset T2.

dataset, both visually and numerically. The accuracy assessment is applied by visual analysis and employing quality measurement indices. To this end, for the purpose of determining the overall validity of the results, we used evaluation metrics such as OA, recall, precision, balanced accuracy (BA), F1-score, Jaccard index/intersection over union (IoU), Kappa coefficient (KC), and Matthews correlation coefficient (MCC). Furthermore, four different state-of-the-art HSI-CD methods were used to compare the effectiveness of the proposed approach: 1) GetNet [24]; 2) TDRD [40]; 3) PTCD [41]; and 4) ViT [42] (standard ViT-based method) that is applied in an unsupervised manner and without any sample dataset. It is worth noting that TDRD and PTCD require threshold selection that uses the K-means algorithm for thresholding. Also, the SSViT model has hyperparameters that need to be set: initial patch size (9 × 9), subdivided size (6 × 6), resampled patch size (24 × 24), weight initializer He-normal manner, learning rate (10^{-3}), number of transformer layers (four), batch size (250), number of transformer layers (four), number of heads (six), dropout rates (0.5 and 0.2), and number of neurons in first and second dense layers (2048 and 1024, respectively). The CBAM mechanism parameters were set as follows: the reduction ratio in 3-D/2-D channel attention module

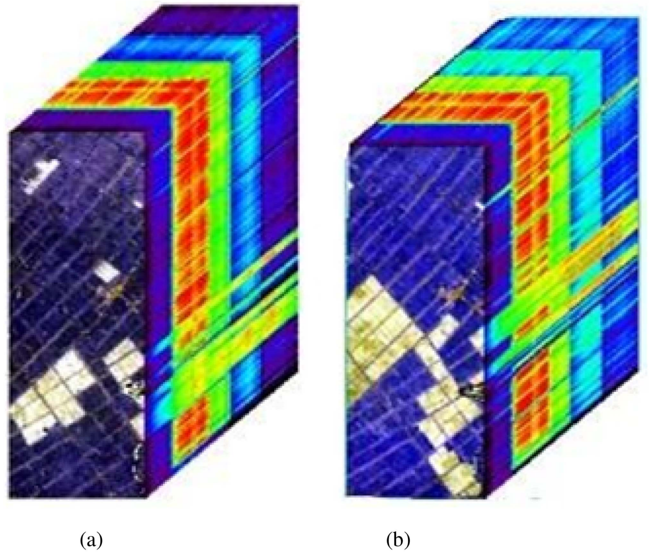


Fig. 10. Farmland dataset. (a) Farmland dataset image T1. (b) Farmland dataset image T2.

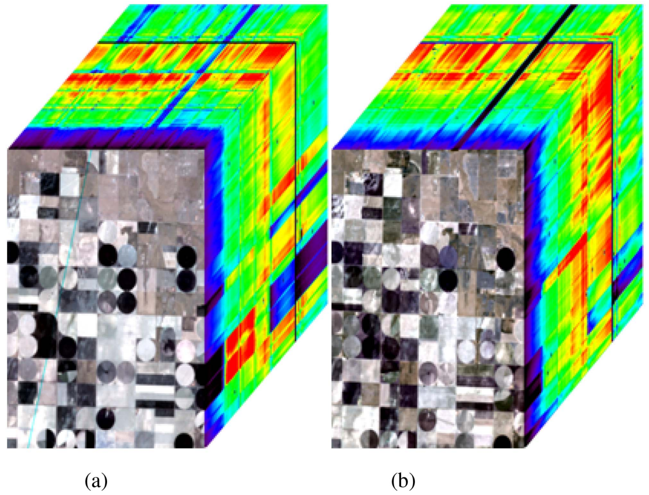
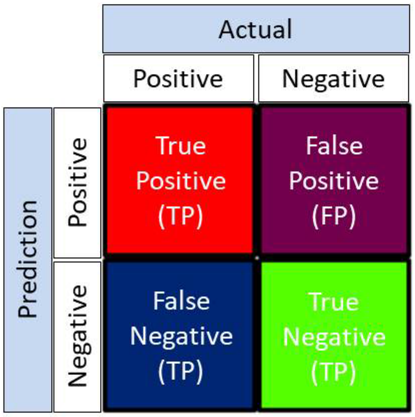


Fig. 11. Agriculture PRISMA dataset. (a) Agriculture PRISMA dataset image T1. (b) Agriculture PRISMA dataset image T2.

is 8 and the kernel sizes for 3-D/2-D spatial attention modules are (3×3) and (7×7), respectively. The Chan–Vese segmentation algorithm includes several parameters: square size $= 5$, edge length parameter $(\mu) = 2$, $\lambda = 0.9$, and a maximum number of iterations $= 25$. These parameters are knowledge-based and set based on trial and error. Moreover, 8000 and 3500 samples were generated for the no-change and change classes, respectively, in the pseudo sample generation.

A. Experimental Results

To evaluate the proposed method thoroughly, we considered the components of the confusion matrix shown in Fig. 12. Several different metrics were carried out as follows.



		Actual	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (TP)	True Negative (TP)

Fig. 12. Components of confusion matrix.

1) *Overall accuracy:* OA can determine immediately whether a model is being trained correctly and how it may perform generally. In the CD applications, due to the imbalance of training data, the use of OA is less effective, so it is necessary to use various criteria for evaluation.
2) *Precision:* The precision measures the proportion of positively predicted labels that are actually correct.
3) *Recall:* When the cost of false negatives is high, recall can indicate the efficiency of the model. In other words, the higher the value of recall, the higher the model's performance in predicting the changed pixels (sacrificing the fact that some of the pixels with the change label may not have changed).
4) *F1-score:* The F1-score combines precision and recall for positive classes, while accuracy considers correctly identified positives and negatives.
5) *Kappa coefficient:* The KC is another metric that measures the reliability of two results (in our methods, GT and model prediction).
6) *Matthews correlation coefficient:* The MCC is a metric that generates a high score if the prediction performed well in all four confusion matrix areas. The MCC also incorporates the dataset imbalance and its invariants for class swapping.
7) *Balanced accuracy:* The problem of the imbalance dataset in the CD application is essential. Therefore, the BA, which is the mean of sensitivity and specificity, is also considered to evaluate the proposed method.
8) Jaccard score (JS)/IoU: The JS obtained from the overlay of the GT and models result, i.e., the ratio of IoU. One of the capabilities of this metric is to show the visual quality of the change map.

In addition to the above quantitative metrics, change maps in the proposed method and comparative methods have been produced. The results are presented separately for each dataset in the following sections.

1) Results on the River Dataset: We first conduct the experiments on a River dataset. The corresponding results of all the quantitative metrics are shown in Fig. 13 and Table I for our proposed SSViT HCD method and four comparison methods. With an overall view at Fig. 13, it can be seen that our proposed SSViT model is superior to other comparative methods in most

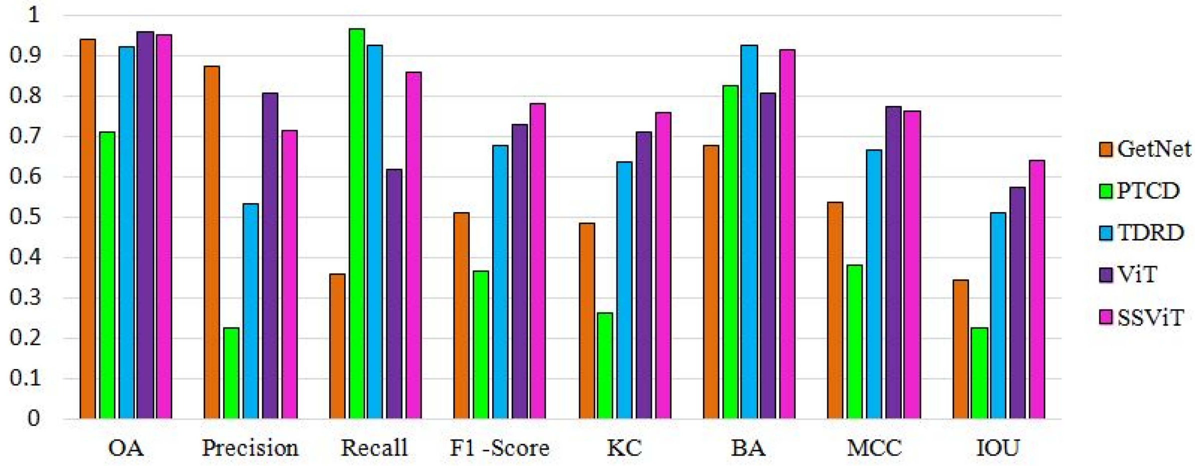


Fig. 13. Quantitative results from the implementation of different methods on the River dataset.

TABLE I
QUANTITATIVE COMPARISONS OF FIVE DIFFERENT METHODS FOR THE DIFFERENT DATASETS

	Index	GetNet	PTCD	TDRD	ViT	SSViT
River dataset	OA (%)	93.98	71.11	92.31	**96.01**	95.80
	Precision (%)	**87.29**	22.69	53.33	80.71	71.55
	Recall (%)	36.05	**96.54**	92.54	61.97	85.90
	F1-Score (%)	51.02	36.74	67.67	72.96	**78.07**
	BA (%)	67.77	82.61	**92.42**	80.61	91.32
	KC	0.484	0.264	0.636	0.709	**0.757**
	MCC	0.538	0.380	0.667	**0.772**	0.761
	IoU	0.342	0.225	0.511	0.574	**0.640**
Hermiston dataset	OA (%)	72.07	96.28	97.20	96.96	**97.54**
	Precision (%)	22.43	92.29	92.01	87.86	**93.87**
	Recall (%)	18.11	88.12	93.64	**97.77**	93.38
	F1-Score (%)	20.04	90.16	92.82	92.55	**93.62**
	BA (%)	51.55	93.18	95.85	**97.27**	95.96
	KC	0.034	0.877	0.911	0.906	0.921
	MCC	0.033	0.879	0.911	0.908	**0.921**
	IoU	0.111	0.821	0.866	0.861	**0.880**
Farmland dataset	OA (%)	**96.11**	95.72	95.28	95.74	95.85
	Precision (%)	**97.43**	89.72	88.20	91.70	91.87
	Recall (%)	89.94	97.50	**98.01**	94.99	95.17
	F1-Score (%)	**93.53**	93.45	92.84	93.32	93.50
	BA (%)	94.43	**96.21**	96.02	95.54	95.67
	KC	**0.907**	0.903	0.893	0.902	0.904
	MCC	**0.909**	0.905	0.896	0.902	0.905
	IoU	0.878	0.877	0.866	0.875	**0.878**
Agriculture PRISMA dataset	OA (%)	88.67	82.56	91.61	94.69	**94.72**
	Precision (%)	87.85	44.96	64.44	87.61	84.44
	Recall (%)	25.67	88.77	**94.52**	80.22	78.14
	F1-Score (%)	39.73	59.70	76.64	**86.09**	81.17
	BA (%)	62.53	85.13	**92.82**	86.09	87.84
	KC	0.355	0.500	71.75	0.772	**0.781**
	MCC	0.438	0.547	0.737	0.775	**0.782**
	IoU	0.248	0.425	0.621	0.670	**0.683**

The best accuracy is in bold in each row.

metrics. The OA of SSViT and ViT is 96%, which is higher than that of all the other methods, showing the higher efficiency of transformation-based methods over classical methods. The highest difference in performance for the proposed method is obtained in KC and F1-score metric, which is about 0.05 more than the ViT, which indicates the reliability of the proposed method. Also, if we divide the compared methods into two parts (classic and transformation-based methods), the difference in the KC reaches 0.12, which is considered a significant difference. Regarding the BA metric, the proposed method and the TDRD have a significant difference from the other two methods, which shows the high accuracy of the changed pixels identified by these two methods. From the point of view of the combined metrics (those metrics consider the true positive and false negative simultaneously), the performance of the five methods follows a similar pattern. Thus, the proposed method, ViT, TDRD, GetNet, and PTCD can be ranked according to performance.

In terms of visual analysis, our proposed method CM quality is highly remarkable, as shown in Fig. 14. From the figures, one can see the impact of various approaches on classification results. When the results are compared to the GT, our proposed SSViT model produces a more accurate CM, confirming that, in addition to quantitative metrics, the visual quality of the change map is superior to other methods. This is due to the inclusion of the pseudo sample generation step in the transformer mechanism.

2) Results on the Hermiston Dataset: Fig. 15 and Table I present the results of the implementation of the proposed method and four comparative methods on the Hermiston dataset. Except for the GetNet method, the efficiency of all other methods is above 0.8 in all metrics, and the value of all metrics is higher than that in the River dataset, which shows that the complexity of Hermiston data is less than that of the River data. The OA of the proposed method is 97.54%. The results obtained from the ViT and TDRD methods are similar to those of the proposed method in most metrics. The highest difference between the methods is obtained in the recall metric, where the proposed method is 0.014 more efficient than the ViT. As shown in the change maps in Fig. 16, it is clear that the quantitative results and the quality of the change map of the GetNet method in these data are significantly low (the amount of false negative is very impressive), which is due to the spectral mixing of this data, which reduces the accuracy of the end members extracted in the initial stage of the GetNet method.

3) Results on the Farmland Dataset: Fig. 17 visualizes the quantitative assessment of the results on the Farmland dataset. All the five methods have performance above 0.87 in all the metrics, as depicted in Table I. The proposed method obtained

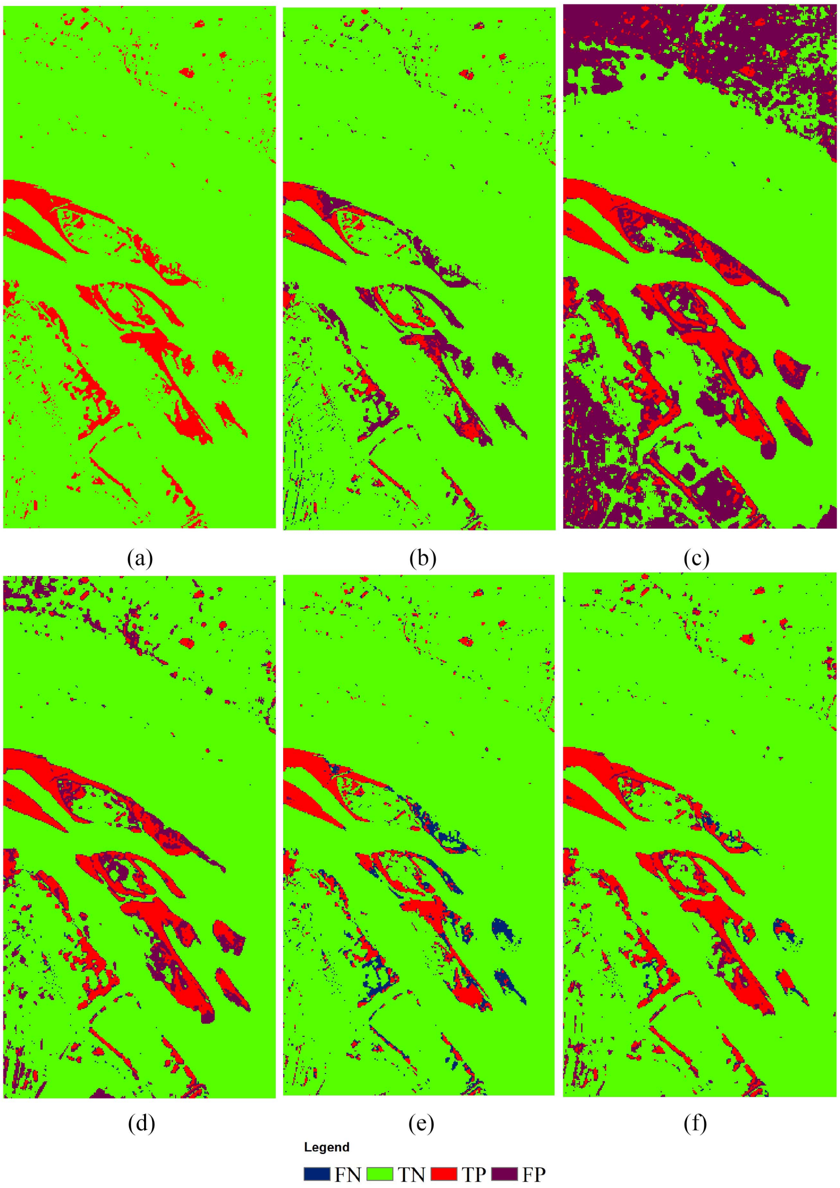


Fig. 14. Change maps obtained by the different methods on the River dataset. (a) Ground truth. (b) Change map obtained by the GetNet method. (c) Change map obtained by the PTCD method. (d) Change map obtained by the TDRD method. (e) Change map obtained by the ViT method. (f) Change map obtained by the proposed method.

the most significant F1-score, KC, and JSs as 0.93, 0.90, and 0.87, as well as GetNet, respectively. Considering the small changes in performance, the transformed-based methods gain the second-best performance in the Farmland dataset after the GetNet method. The visual quality of the change map shown in Fig. 18 indicates that the performance of all methods is very similar.

4) Results on the Agriculture PRISMA Dataset: The implementation results on the Agriculture PRISMA dataset show a significant difference between the methods, as shown in Table I and Fig. 19. This dataset can well validate our proposed method to its advantages in eliminating the problem of a lack of quality and quantity of training samples. According to the results of this dataset, some conclusions are similar to those obtained by the Hermiston dataset. The proposed SSViT method achieves the highest quantitative assessment performance with the OA of 0.95, the KC of 0.78, the MCC of 0.78, and the F1-score of 0.86.

Fig. 20 presents the visual output of all the methods. By comparing the results of this dataset and, to some extent other datasets, it can be seen that the weakness and strengths of the methods are at the edge of the areas that have changed. Our proposed method has performed better than other methods in the isolated regions that have changed, unlike the other methods, which mainly have a high false negative, which has reduced their accuracy. This difference shows that the training data that are the input of the proposed model are of high quality.

5) Ablation Analysis: In artificial intelligence methods, the ablation analysis is a crucial step in getting insight into the overall performance by removing parts of the system [43]. In four scenarios, the study examines the influence of the attention

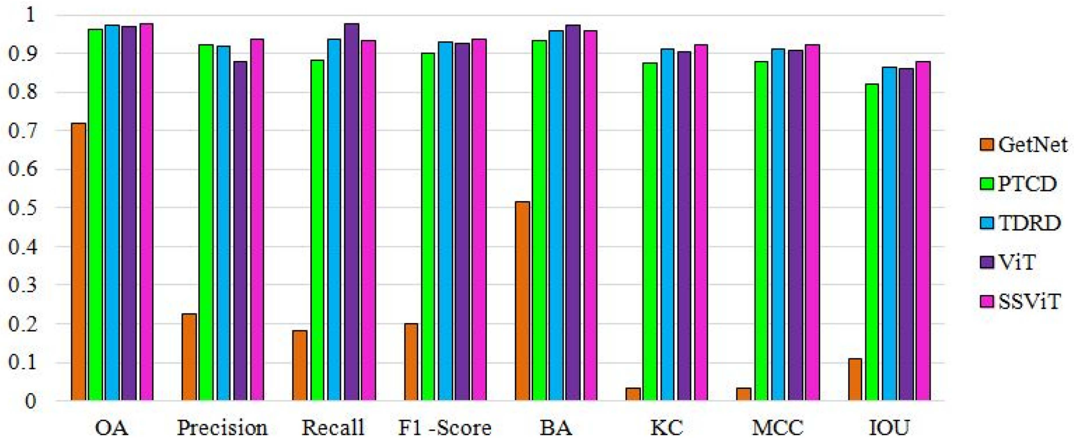
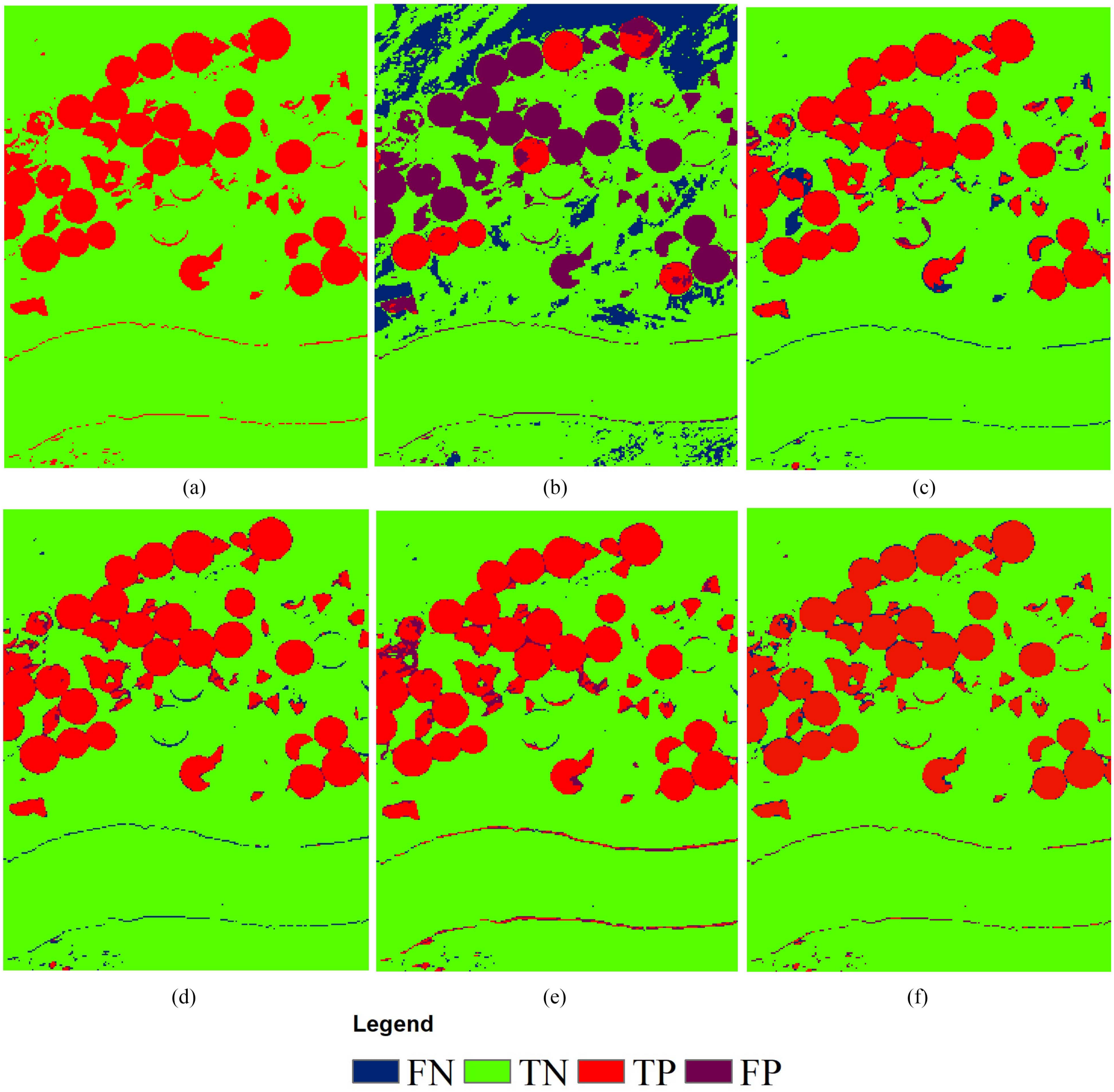


Fig. 15. Quantitative results from the implementation of different methods on the Hermiston dataset.



(a) (b) (c)

(d) (e) (f)

Fig. 16. Change maps obtained by the different methods on the Hermiston dataset. (a) Ground truth. (b) Change map obtained by the GetNet method. (c) Change map obtained by the PTCD method. (d) Change map obtained by the TDRD method. (e) Change map obtained by the ViT method. (f) Change map obtained by the proposed method.

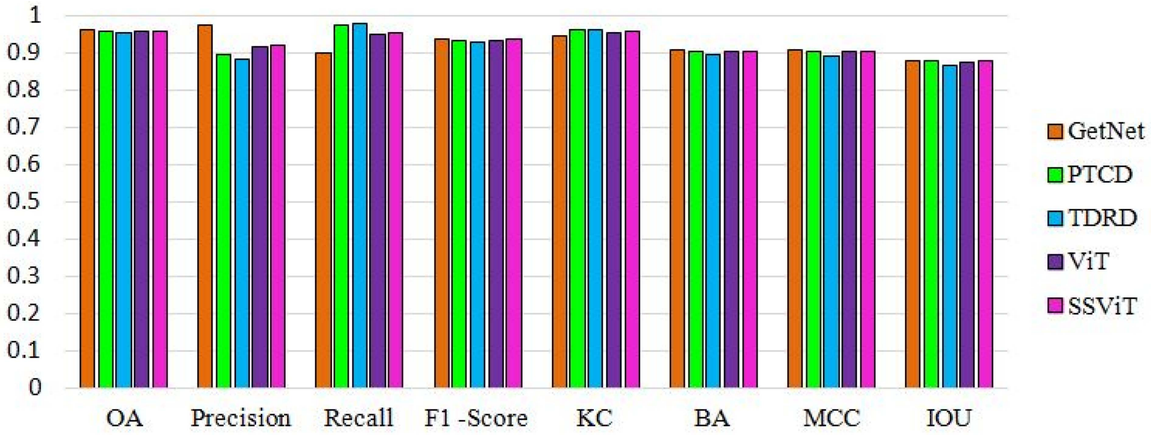


Fig. 17. Quantitative results from the implementation of different methods on the Farmland dataset.

TABLE II
Ablation Analysis: Quantitative Comparisons Among Other Variants of the Proposed SSViT Algorithm for the Hermiston Dataset

Index	**S(#1)**	**S(#2)**	**S(#3)**	**S(#4)**	**SSViT**
OA (%)	97.24	97.01	96.61	96.96	**97.54**
Precision (%)	93.01	95.33	90.95	87.86	**93.86**
Recall (%)	92.71	88.90	91.61	**97.77**	93.38
F1-Score (%)	92.86	92.01	91.27	92.55	**93.62**
BA (%)	95.52	93.93	94.71	**97.27**	95.96
KC	0.911	0.901	0.892	0.906	**0.921**
MCC	0.911	0.902	**0.982**	0.908	0.921
IoU	0.866	0.852	0.839	0.861	**0.880**

The bold font shows the highest value.

TABLE III
Comparison of Computational Cost of DL-Based Methods

Models	**GetNet**	**ViT**	**SSViT**
Number of parameters	24 (M)	2.9 (M)	{2.2 (M)

module, transformer module, and convolution parts on the proposed methods' performance. The first scenario (S#1) is to remove the 2-D attention module, the second scenario (S#2) is to remove the 3-D attention module, the third scenario (S#3) is to remove the transformer part, and the fourth scenario (S#4) is to remove all multiscale convolution layers and attention modules. The numerical experimental results of ablation analysis for the Hermiston dataset are presented in Table II. As seen, all parts' effectiveness is different from each other. Removing the transformer module clearly has the highest negative impact on the model's performance in HCD. Furthermore, the 2-D CBAM has the lowest influence on the model's efficacy. However, removing all multiscale convolution layers and attention modules (S#4) provides better performance by recall and BA indices, but it misses its performance by other metrics.

6) Computational Cost: We evaluated the computational cost of HCD methods. However, the PTCD and TDRD are not DL-based frameworks; we compared the three DL-based models, such as GetNet, ViT, and SSViT. Table III shows the number of parameters of models and that SSViT has lower number of parameters than that of the other two DL-based models. It is worth noting that the lower parameters can improve the time processing and computational cost.

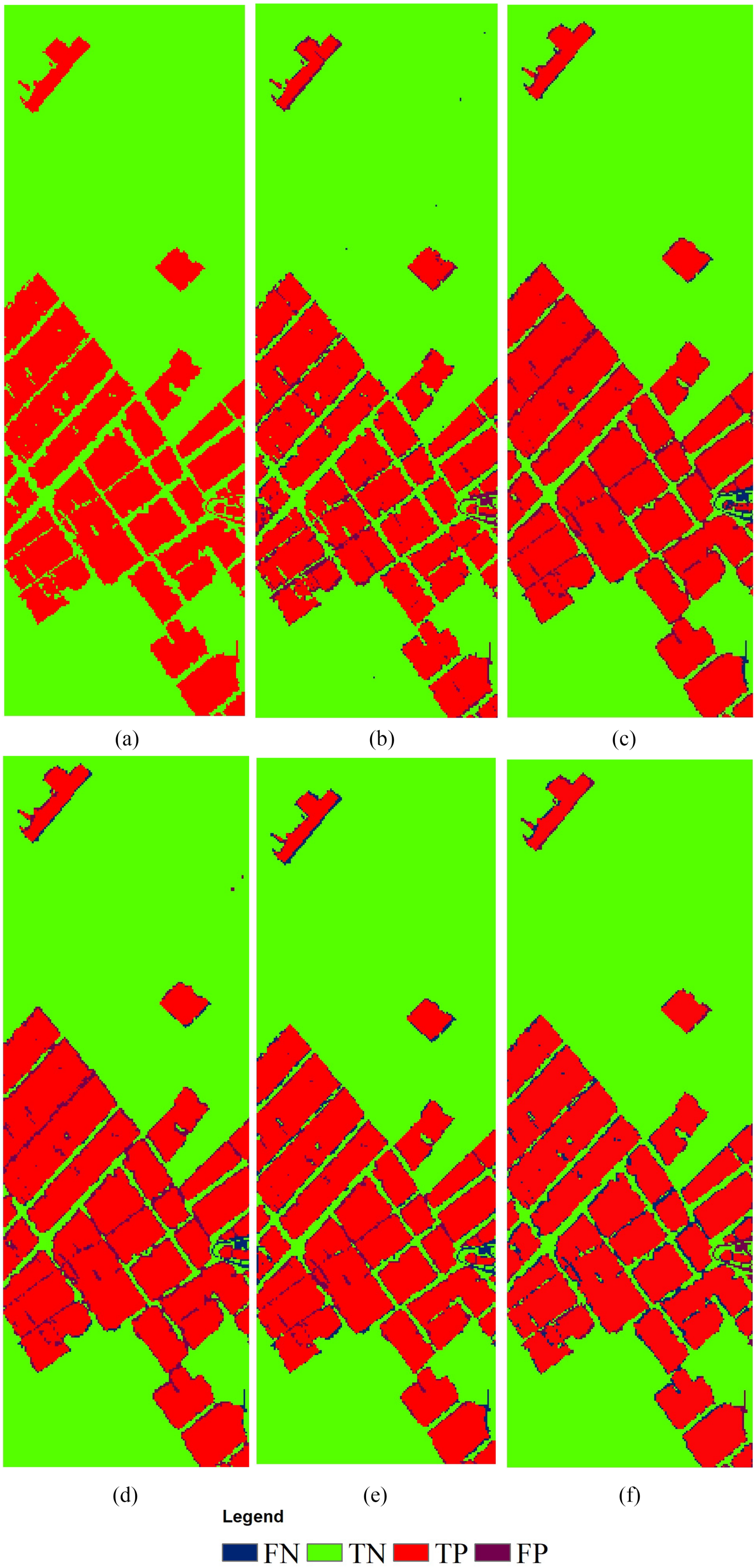
V. Discussion

We present a novel and efficient method to perform HCD known as the SSViT, which combines the benefits of the CNN and the transformer model. Our experiment was conducted on four HSI datasets, and we compared the experimental results obtained using SSVit with the experimental results obtained using the other four hyperspectral-based CD models. According to Table I, in terms of most metrics, the proposed SSViT model has a significant advantage over other comparative methods from the overall performance perspective. One important aspect is worth mentioning: the performance of SSViT is high, which indicates that SSViT has apparent advantages over classical methods in HCD.

Our proposed method generates a more accurate and complete change map that is close to the GT. In addition to quantitative metrics, the visual quality of the change map is much better than that of other methods, as shown in Figs. 14 and 16. Briefly, all the visual interpretations of the CMs in the Fig. 20 qualitatively reflect the effectiveness of the proposed SSViT model.

Furthermore, there is tradeoff between change and no-change pixels in HCD. Some models only consider the change pixels, while ignoring its performance in the no-change pixels. Inversely, some models only focused on no-change, while others missed their performance in the change pixels or had a challenge with a large number of mixed pixels, such as [32]. All the models should be able to identify change and no-change pixels that have the lowest error and can effectively improve the CD ability. For example, from Table I, in the River dataset, the GetNet algorithm has provided high precision while missing performance in recall under 36%. In addition, in the Farmland dataset, the TDRD has provided high accuracy by recall but missed performance in precision under 88%. F1-score originates from recall and precision, and the proposed SSViT model has provided high value in more datasets, which means that the proposed model focuses more on change and no-change pixels because it provides high value by F1-score.

In addition, to demonstrate the significance of the ViT model and the CNN model in capturing local and global features in Fig. 21, we present the visualization of feature maps of the



(a) (b) (c)

(d) (e) (f)

Fig. 18. Change maps obtained by the different methods on the Farmland dataset. (a) Ground truth. (b) Change map obtained by the GetNet method. (c) Change map obtained by the PTCD method. (d) Change map obtained by the TDRD method. (e) Change map obtained by the ViT method. (f) Change map obtained by the proposed method.

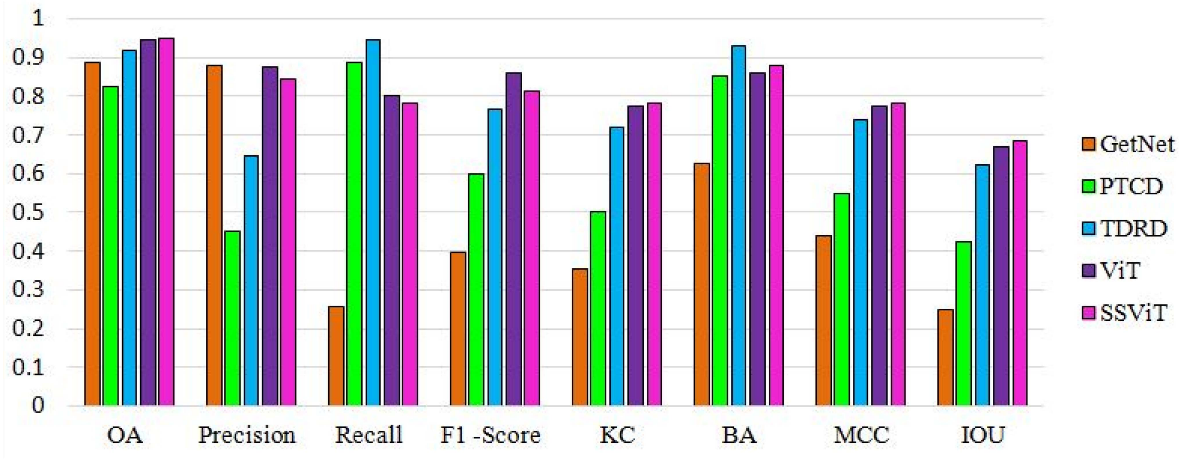


Fig. 19. Quantitative results from the implementation of different methods on the Agriculture PRISMA dataset.

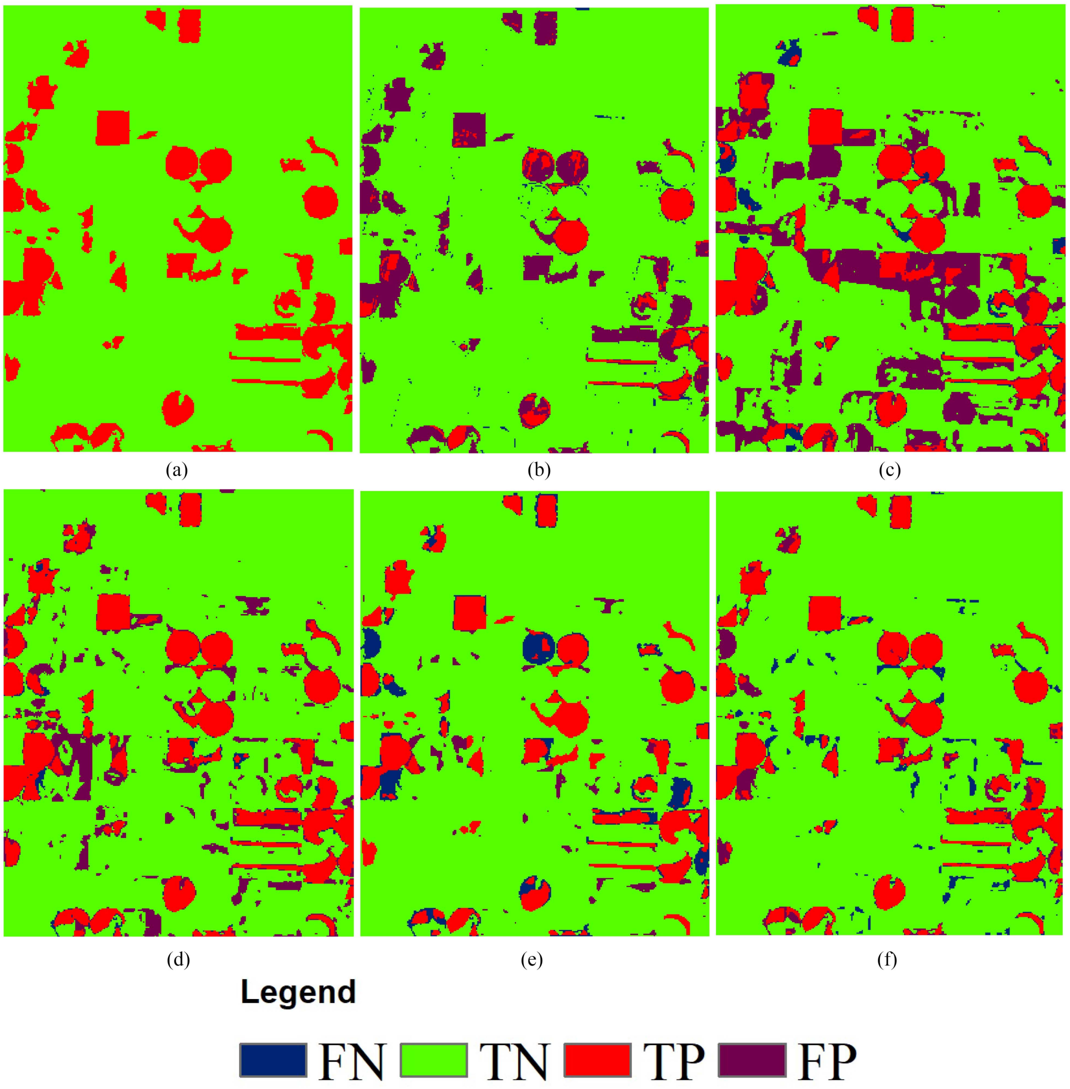


Fig. 20. Change maps obtained by the different methods on the Agriculture PRISMA dataset. (a) Ground truth. (b) Change map obtained by the GetNet method. (c) Change map obtained by the PTCD method. (d) Change map obtained by the TDRD method. (e) Change map obtained by the ViT method. (f) Change map obtained by the proposed method.

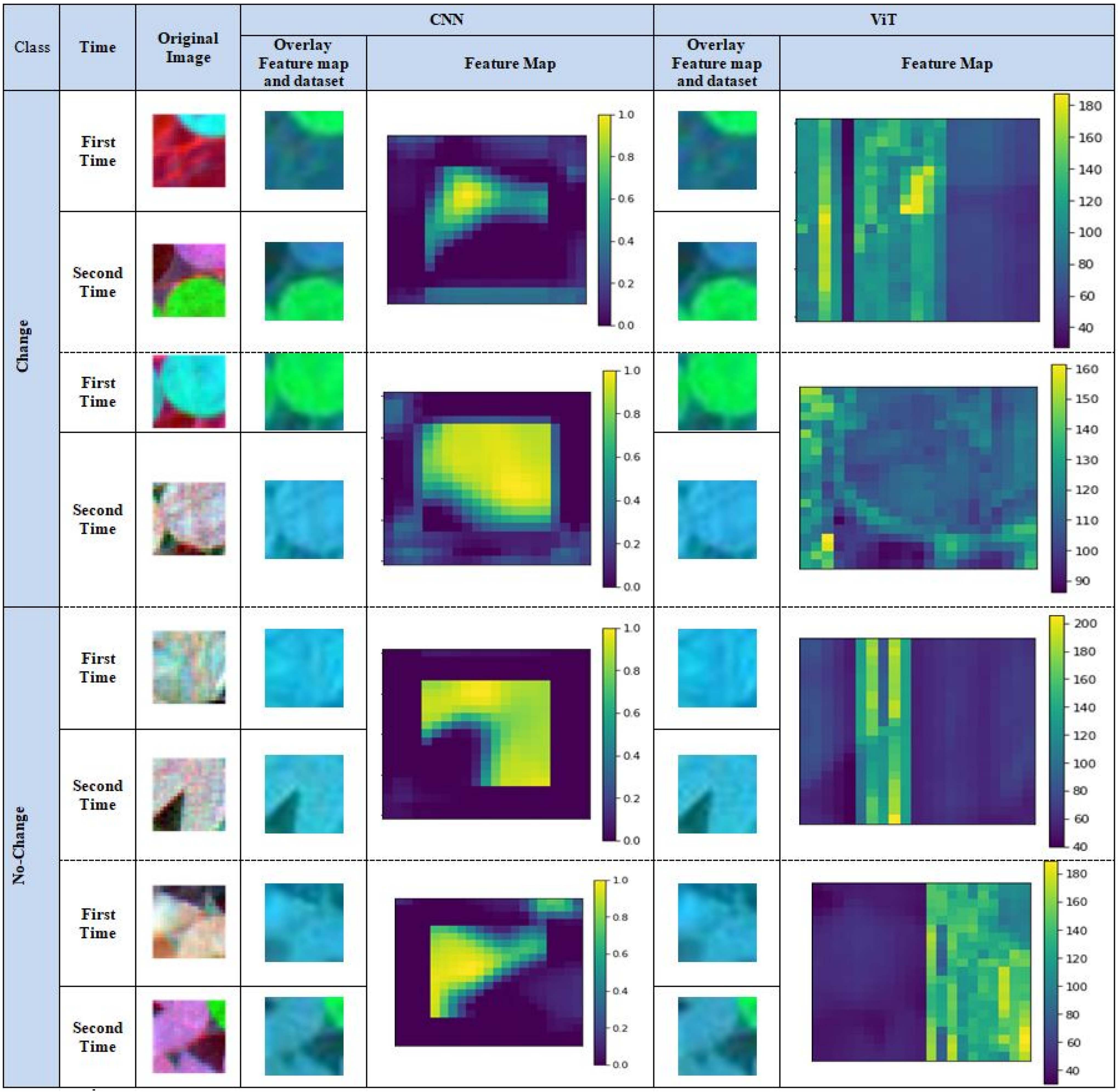


Fig. 21. Feature map comparison between CNN and ViT models.

CNN and the ViT model. The convolution layer extracts the local feature, while the ViT model can extract the global feature. Fig. 21 shows the effectiveness of the ViT model in capturing the long-range modeling relationships in the input patches. Furthermore, the result of the visualization of both the models shows that the CNN model is focused on local features because it captures local information on the feature map, and there is less dependence among long-range dependencies in the feature maps (yellow areas in the feature map), as shown in Fig. 21. The transformer models are specialized for long-term relationships, as there are many yellow points in the different areas of the feature maps. Our proposed method, SSViT, combines the benefits of the convolution layer and the transformer model and can separately extract local and global features, which improves the result of HCD.

DL models require a massive amount of samples, which can challenge bitemporal datasets [44], [45]. Also, the quality and size of the sample dataset are still some of the problems that supervised methods have to deal with. Moreover, the proposed method did not require the collection of user training data or the setting of parameters. Several HCD approaches have recently been presented to generate sample data using an unsupervised framework to enhance the sample data's reliability [10], [24]. These approaches primarily create sample data using classic predictors (e.g., PCA and EU predictors). However, they find satisfactory results, but errors in sample generation can affect the final CD. Therefore, reliable sample generation is still challenging. Due to unreliable sample data, supervised classifiers are trained with false data, resulting in low accuracy. A significant achievement of the proposed method was refining sample data

TABLE IV
COMPARISON OF ACCURACIES OF THE SSViT WITH OTHER HCD METHODS

Methods	**Datasets**	**OA**	**Reference**
MCS^{4}CD	River	93.46%	[46]
Subpixel	Farmland	92.32%	[47]
D^{2}AGCN	Farmland	0.9374%	[34]
SSCNN-S	Hermiston	0.9651%	[48]
MCS^{4}CD	Hermiston	94.46%	[46]
Proposed method	River	95.80%	—
Proposed method	Farmland	95.85%	—
Proposed method	Hermiston	97.54%	—

using a hierarchical thresholding method to make sample data more reliable. Furthermore, Table III shows the number of parameters of two comparison methods and our proposed method. We found that SSViT has fewer parameters (2.2 M) than other DL-based models. The number of parameters is high in other methods; it means that they take a lot of time for processing and have a high computational cost.

To further demonstrate the superiority of the model, we collected the accuracy of the datasets presented in their original work and listed them in Table IV. The proposed model in [46] acquires the OA of 93.46% on the River dataset. The Farmland dataset was used in [47], which had an accuracy of 92.32%, and in [34], which achieved an accuracy of 93.74%. Furthermore, an OA of 96% was obtained with the Hermiston dataset [48], and the model in [46] achieved an OA of 94.46%. Our proposed method, SSViT, obtained the highest accuracies for similar datasets, such as River, Farmland, and Hermiston, as shown in Table IV. Also, it is important to note that we did not compare the accuracy of the PRISMA dataset with the previous DL-based CD method because we first used the PRISMA dataset for HCD.

VI. CONCLUSION

In this article, we proposed a new hybrid SSViT framework for HCD. Our proposed model combines the merit of CNN and transformer. In general, the proposed technique provides several advantages over other CD methods: 1) it provides an end-to-end framework; 2) it generates a reliable sample dataset that directs the CD for automatic framework; 3) it is robust and high efficient in different complex land cover areas; and 4) it is adaptive with a different hyperspectral dataset captured by different sensors. Moreover, this article provided a novel means for hyperspectral-based CD. Therefore, this study proved that SSViT provides significant results in the HCD task. Furthermore, new series of HSIs was introduced for CD purposes. In the future, we will keep researching transformer use in RS images for CD tasks.

REFERENCES

[1] H. Shi, G. Cao, Z. Ge, Y. Zhang, and P. Fu, "Double-branch network with pyramidal convolution and iterative attention for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 7, 2021, Art. no. 1403.

[2] Y. Zhang, G. Cao, A. Shafique, and P. Fu, "Label propagation ensemble for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3623–3636, Sep. 2019.

[3] A. Moghimi, A. Sarmadian, A. Mohammadzadeh, T. Celik, M. Amani, and H. Kusetogullari, "Distortion robust relative radiometric normalization of multitemporal and multisensor remote sensing images using image features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5400820.

[4] W. A. Khanday and K. Kumar, "Change detection in hyper spectral images," *Asian J. Technol. Manage. Res.*, vol. 6, no. 2, pp. 54–60, 2016.

[5] X. He and Y. Chen, "Optimized input for CNN-based hyperspectral image classification using spatial transformer network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1884–1888, Dec. 2019.

[6] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715.

[7] Y. Yuan, D. Ma, and Q. Wang, "Hyperspectral anomaly detection by graph pixel selection," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3123–3134, Dec. 2016.

[8] N. Ma, Y. Peng, S. Wang, and P. H. Leong, "An unsupervised deep hyperspectral anomaly detector," *Sensors*, vol. 18, no. 3, 2018, Art. no. 693.

[9] G. Cao, B. Wang, H.-C. Xavier, D. Yang, and J. Southworth, "A new difference image creation method based on deep neural networks for change detection in remote-sensing images," *Int. J. Remote Sens.*, vol. 38, no. 23, pp. 7161–7175, 2017.

[10] S. T. Seydi, R. Shah-Hosseini, and M. Amani, "A multi-dimensional deep siamese network for land cover change detection in bi-temporal hyperspectral imagery," *Sustainability*, vol. 14, no. 19, 2022, Art. no. 12597.

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

[12] 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, 2020, Art. no. 2010.

[13] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with landsat," in *Proc. LARS symposia*, 1980, p. 385.

[14] H. Zhuang, K. Deng, H. Fan, and M. Yu, "Strategies combining spectral angle mapper and change vector analysis to unsupervised change detection in multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 681–685, May 2016.

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

[16] S. Singh and R. Talwar, "A comparative study on change vector analysis based change detection techniques," *Sadhana*, vol. 39, no. 6, pp. 1311–1331, 2014.

[17] V. Ortiz-Rivera, M. Vélez-Reyes, and B. Roysam, "Change detection in hyperspectral imagery using temporal principal components," *Proc. SPIE*, vol. 6233,, pp. 368–377, 2006.

[18] 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.

[19] 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.

[20] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[21] L. Wiskott, P. Berkes, M. Franzius, H. Sprekeler, and N. Wilbert, "Slow feature analysis," *Scholarpedia*, vol. 6, no. 4, 2011, Art. no. 5282.

[22] C. Wu, L. Zhang, and B. Du, "Kernel slow feature analysis for scene change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2367–2384, Apr. 2017.

[23] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogrammetry Remote Sens.*, vol. 80, pp. 91–106, 2013.

[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] Y. Yuan, H. Lv, and X. Lu, "Semi-supervised change detection method for multi-temporal hyperspectral images," *Neurocomputing*, vol. 148, pp. 363–375, 2015.

[26] S. Liu, L. Bruzzone, F. Bovolo, and P. Du, "Unsupervised multitemporal spectral unmixing for detecting multiple changes in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2733–2748, May 2016.

[27] A. Ertürk, M.-D. Iordache, and A. Plaza, "Sparse unmixing with dictionary pruning for hyperspectral change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 1, pp. 321–330, Jan. 2017.

[28] C. Wu, B. Du, and L. Zhang, "Hyperspectral anomalous change detection based on joint sparse representation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 146, pp. 137–150, 2018.

[29] X. Li, Z. Yuan, and Q. Wang, "Unsupervised deep noise modeling for hyperspectral image change detection," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 258.

[30] 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, 2018, Art. no. 1827.

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

[32] F. Huang, Y. Yu, and T. Feng, "Hyperspectral remote sensing image change detection based on tensor and deep learning," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 233–244, 2019.

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

[34] J. Qu, Y. Xu, W. Dong, Y. Li, and Q. Du, "Dual-branch difference amplification graph convolutional network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5519912.

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

[36] P. Getreuer, "Chan-Vese segmentation," *Image Process. Line*, vol. 2, pp. 214–224, 2012.

[37] R. Cohen, "The Chan-Vese algorithm project report," Technion—Israel Inst. Technol., Haifa, Israel, Tech. Rep., 2010.

[38] R. Cohen, "The Chan-Vese algorithm," 2011, *arXiv:1107.2782*.

[39] A. Munir, S. Soomro, C. H. Lee, and K. N. Choi, "Adaptive active contours based on variable Kernel with constant initialisation," *IET Image Process.*, vol. 12, no. 7, pp. 1117–1123, 2018.

[40] 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.

[41] Z. Hou, W. Li, and Q. Du, "A patch tensor-based change detection method for hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4328–4331.

[42] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[43] S. T. Seydi, M. Amani, and A. Ghorbanian, "A dual attention convolutional neural network for crop classification using time-series sentinel-2 imagery," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 498.

[44] S. T. Seydi, M. Hasanlou, and J. Chanussot, "Burnt-Net: Wildfire burned area mapping with single post-fire Sentinel-2 data and deep learning morphological neural network," *Ecol. Indicators*, vol. 140, 2022, Art. no. 108999.

[45] S. T. Seydi, M. Hasanlou, and J. Chanussot, "A quadratic morphological deep neural network fusing radar and optical data for the mapping of burned areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4194–4216, 2022.

[46] L. Liu, D. Hong, L. Ni, and L. Gao, "Multilayer cascade screening strategy for semi-supervised change detection in hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1926–1940, 2022.

[47] M. Hasanlou, S. T. Seydi, and R. Shah-Hosseini, "A sub-pixel multiple change detection approach for hyperspectral imagery," *Can. J. Remote Sens.*, vol. 44, no. 6, pp. 601–615, 2018.

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

Ayesha Shafique received the M.S. degree in computer science from Government College University, Faisalabad, Pakistan, in 2018. She is working toward the Ph.D. degree in computer science and technology with the Nanjing University of Science and Technology, Nanjing, China.

She is an active reviewer of some remote sensing journals. Her research interests include image processing, machine learning, deep learning, and change detection in remote sensing images.

Seyd Teymoor Seydi received the B.Eng. degree in surveying and geomatics engineering from the University of Shahid Rajaei, Tehran, Iran, in 2015, and the M.Eng. degree in remote sensing from the University of Tehran, Tehran, in 2018.

He is currently with the University of Tehran. He serves as a regular reviewer for about 14 international remote sensing journals. He authored or coauthored more than 40 peer-reviewed journal and conference papers. His research interests include multitemporal multispectral/hyperspectral and synthetic aperture radar remote sensing processing and classification, and advanced deep learning algorithms.

Tayeb Alipour-Fard received the Ph.D. degree in remote sensing and photogrammetry from the School of Surveying and Geospatial Engineering, University of Tehran, Tehran, Iran, in 2021.

He was a Visiting Researcher with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. His research interests include machine learning, pattern recognition, deep learning, feature extraction, and classification of hyperspectral imagery.

Dr. Alipour-Fard has been a Reviewer for IEEE Geoscience and Remote Sensing Letters and IEEE Transactions on Geoscience and Remote Sensing since 2018

Guo Cao received the Ph.D. degree in pattern recognition and intelligence system from Shanghai Jiao Tong University, Shanghai, China, in 2006.

Since 2007, he has been with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, where he is currently a Full Professor. From 2012 to 2013, he was a Visiting Scholar with the Department of Radiology, University of Chicago, Chicago, IL, USA. From 2016 to 2017, he was a Visiting Scholar with the Department of Geography, University of Florida, Gainesville, FL, USA, where he focused on change detection. His research interests include machine learning, remote sensing image processing, and biometrics.

Di Yang received the Ph.D. degree in geography from the University of Florida, Gainesville, FL, USA, in 2019.

She is currently an Assistant Professor with the Wyoming Geographic Information Science Center, University of Wyoming, Laramie, WY, USA. She is a Geographer and Geospatial Data Scientist with expertise in geospatial informatics, applied remote sensing, volunteered geographic information, and macrosystem ecology. She is an active reviewer of many remote sensing journals. Her research interests include the development of advanced remote sensing techniques through digital image processing and geovisualization applications, specifically in coupled human–environment interactions and citizen sciences. She is also interested in developing and sharing expertise in processing and analyzing large ecological and remote sensing database by using cutting edge tools (e.g., Google Earth Engine, Microsoft Azure, and Web-GIS) to benefit conservation, resource management, landowners, and policy makers.