A Spectrum-Aware Transformer Network for Change Detection in Hyperspectral Imagery

Wuxia Zhang, Liangxu Su, Yuhang Zhang, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*— Change detection in the hyperspectral imagery (HSI) detects the changed pixels or areas in bitemporal images. HSIs contain hundreds of spectral bands, including a large amount of spectral information. However, most of deep learning-based change detection methods did not focus on the spectral dependency of spectral information in the spectral dimension and just adopted the difference strategy to represent the correlation of learned features, which limited the improvement of the change detection performance. To address the above-mentioned problems, we propose an end-to-end change detection network for HSIs, named spectrum-aware transformer network (SATNet), which includes SETrans feature extraction module, the transformer-based correlation representation module, and the detection module. First, SETrans feature extraction module is employed to extract deep features of HSIs. Then, the transformer-based correlation representation module is presented to explore the spectral dependency of spectral information and capture the correlation of learned features of bitemporal HSIs from both the perspective of difference and dot-product operations, so as to obtain more discriminative features. Finally, the decision fusion strategy in the detection module is utilized to the learned discriminative features to generate the final change map for better change detection performance. Experimental results on three hyperspectral datasets show that the proposed SATNet is superior to the existing change detection methods.**

***Index Terms*— Change detection, deep learning, hyperspectral images, transformer.**

I. Introduction

CHANGE detection of remote sensing images (RSIs) aims to detect changes of bitemporal RSIs at different times, and then make necessary analysis of changed regions from the quantitative and qualitative perspectives. It has been widely used in many fields: city management [1], [2], nature disaster monitor [3], [4], forest management [5], [6], and so on. Hyperspectral image (his) is a typical RSI with hundreds or even thousands of spectral bands. These images contain rich spectral and detailed information, which can be used to detect or identify targets. Hence, hyperspectral change detection attracts more and more attention and has become a hot topic in the field of RSI processing.

Change detection methods of RSIs can be roughly classified into traditional change detection methods and deep learning-based methods. The traditional change detection methods mainly include algebra-based methods (such as image difference [7], image rationing [8]), transformation-based methods (such as change vector analysis (CVA) [9], principal component analysis (PCA) [10]), classification-based method [11], and other methods (Markov model [12], fuzzy clustering [13]). Most traditional change detection methods rely on manual features or shallow features that cannot represent targets very well, which leads to the low-change detection accuracy.

Deep learning techniques have been widely used in the field of computer vision due to their ability to extract abstract, hierarchical, and high-level features. Hence, deep learning techniques have already been applied to the remote sensing change detection task. Various deep learning structures have been utilized to process HSIs, such as feature pyramid network (FPN) [14], [15], Siamese network [1], [16], [17], Unet [18], [19], and so on. However, these change detection methods do not pay attention to the regions of interest (changed areas), and the features extracted from these regions are more important to improve the detection performance. Hence, attention-based change detection methods [20], [21] are presented to address this problem. They can focus on the regions of interest and extract more representative features through the generated spatial or channel attention maps. Long-range dependencies have been proven to improve the detection performance. Therefore, the transformer has been introduced to process RSIs since it can easily focus on long-range dependencies. Transformer-based change detection methods [22], [23] [24], [25] [26], [27] can interact with long-distance information, and capture the relationship between HSIs to obtain discriminative features.

Although the above-mentioned transformer-based change detection methods have already achieved good performance in the change detection task, most of these methods only introduced the transformer mechanism to explore the position dependency of spatial information in the spatial dimension and did not employ the transformer mechanism to mine the spectral dependency of spectral information in the spectral dimension. However, due to a large number of spectral bands in HSIs, the spectral information of HSIs is more important for detecting changes. Moreover, most deep learning-based change

Manuscript received 19 November 2022; revised 7 March 2023 and 19 June 2023; accepted 15 July 2023. Date of publication 28 July 2023; date of current version 11 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62001378 and in part by the Shaanxi Province Network Data Analysis and Intelligent Processing Key Laboratory Open Fund under Grant XUPT-KLND(201902). *(Corresponding author: Wuxia Zhang.)*

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

Liangxu Su is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

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

Digital Object Identifier 10.1109/TGRS.2023.3299642

detection methods only performed the difference operation to calculate the correlation between learned deep features and did not explore the correlation between the bitemporal images from other perspectives.

To address the above-mentioned two problems, we propose an end-to-end change detection network, named spectrum-aware transformer network (SATNet). The proposed SATNet includes SETrans feature extraction module, the transformer-based correlation representation module, and the detection module. The SETrans feature extraction module uses a series of squeeze units and expand units to extract the deep spatial features and fuses the transformer encoder to interact the spectral information. The transformer-based correlation representation module extracts the correlation of his information at different times from two standpoints of difference and dot-product operations. The transformer encoder is employed in both branches to fuse the spectral information in each band to obtain the correlation of the spectral features. Finally, the detection module presents a novel decision fusion strategy to make full use of learned features by weighted summing the obtained change maps, which can improve the accuracy and robustness of change detection.

The main contributions of our article are summarized as follows.

1) We propose a novel SATNet fhisHSI change detection, which uses the transformer mechanism in the spectral dimension to explore the spectral dependency of spectrum.
2) We design a transformer-based correlation representation module to capture the correlation of learned features from the SETrans feature extraction module from the perspective of differential and multiplicative streams, so as to extract more discriminative features.
3) We present a new decision fusion strategy to obtain a more accurate change detection map by weighted integrating the detection results of the differential stream, multiplicative stream, and their corresponding concatenating stream.

The rest of the article is structured as follows. Section II reviews the related literatures of change detection methods. Section III describes the proposed SATNet in detail. The effectiveness of the proposed SATNet is verified on three real datasets in Section IV. Finally, Section V is the conclusion.

II. Related Work

A. Deep Learning-Based Change Detection Method

Deep learning-based change detection methods have become the mainstream for the change detection task due to their strong feature representation. Deep learning-based change detection methods can be roughly classified into two groups: supervised learning-based methods and unsupervised learning-based methods.

Supervised learning-based change detection methods [17], [28], [29], [30], [31], [32] refer to training a network with labeled data and transforming the change detection task into a classification problem through the design of loss functions. Yang et al. [17] proposed a transferred deep learning-based change detection framework, which reduced the distribution discrepancy between source and target domains. Many network structures are used or improved in the change detection task with the supervised learning strategy. Zhang et al. [28] independently trained a fully convolutional two-stream architecture to extract features from the high-resolution bitemporal RSIs. Seydi et al. [29] performed a multidimensional convolution structure to extract deep features, which includes three types of convolution kernels: 1-D-, 2-D-, and 3-D-dilated-convolution. Li et al. [30] used multiscale convolution kernels to extract the detailed features of land covers. Zhang et al. [31] presented a supervised method based on deep Siamese semantic network, which is trained by the improved triplet loss for optical aerial images. Wiratama et al. [32] combined dense connection convolution layers into the dual network to reuse preceding feature maps and better measure the dissimilarity of two temporal images.

Since the supervised learning-based change detection method requires a large amount of labeled data, it is time-consuming and laborious to tag RSIs. To address this problem, many semi-supervised and unsupervised learning-based change detection methods [33], [34], [35], [36], [37] have attracted more and more attention. Yu et al. [33] developed a band selection approach to choose a salient subset implied sufficient information, and a low-rank representation model and a cluster algorithm are used to acquire the change detection map. Yu et al. [34] introduced the unsupervised domain adaptation strategy with dense feature compaction to migrate the model trained in the labeled samples of the source domain to the unlabeled target domain. Sadeghi et al. [35] proposed an unsupervised fuzzy measurement approach for multitemporal and multispectral RSIs based on the asymmetric thresholding and fuzzy logic. Li et al. [36] presented a noise modeling-based fully convolutional network for HSI change detection, which is trained by the existing unsupervised change detection methods and then removed the noise during the training process.

B. Attention-Based Change Detection Method

Attention mechanisms have been widely used in the field of computer vision to emphasize the feature representations of important areas while suppressing features of unnecessary areas. This is achieved by learning the weight distribution of features through generating a mask that highlights the key parts of an image. The main difference between the attention mechanism and the receptive field of convolutional neural networks (ConvNets) is that the attention mechanism is typically used to selectively weight different regions of an image or feature map, rather than simply adjusting the receptive field of individual neurons. Attention-based change detection methods for RSIs can be divided into three categories: spatial attention-based change detection method [20], channel attention-based change detection method [21], and spatial-channel attention-based change detection method [38], [39], [40], [41], [42].

Chen and Shi [20] proposes a spatial–temporal attention-based change detection method for RSIs, which partitioned the image into multiscale subregions and introduced a

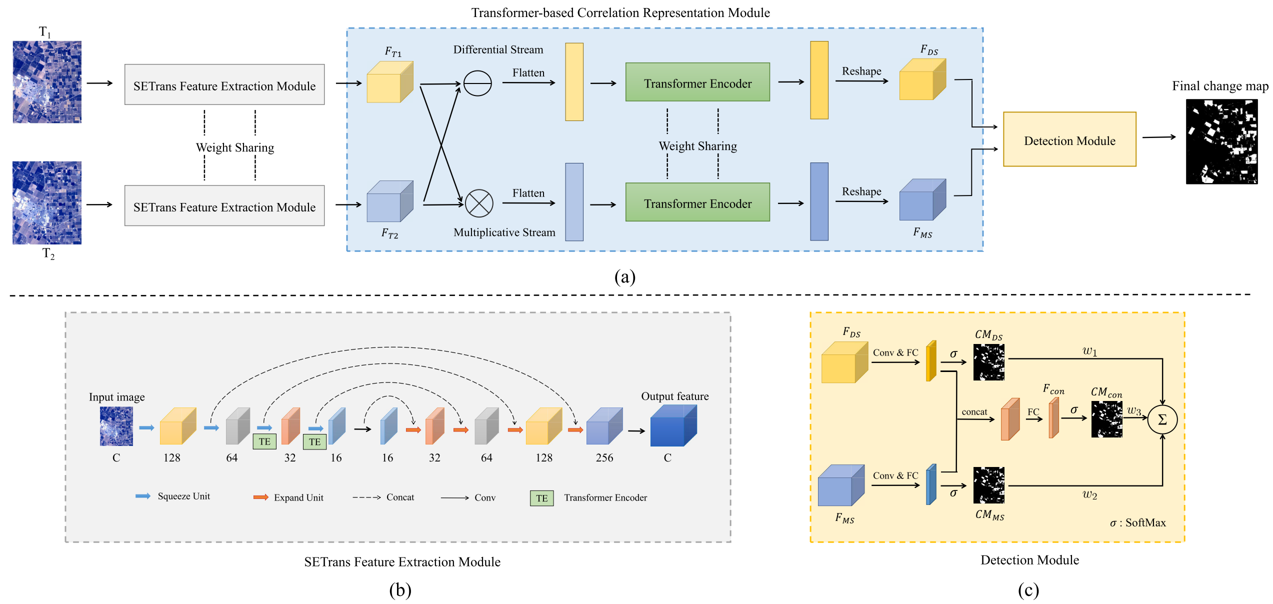


Fig. 1. Structure of the SATNet. (a) Framework of the SATNet. (b) Structure of SETrans feature extraction module and the corresponding number below each feature is the spectral dimension. (c) Structure of the detection module.

spatial–temporal attention module to capture spatial–temporal dependencies at different scales. Peng et al. [21] presented a semi-supervised CNN for change detection, which introduced the channel attention mechanism to refine the Unet++ model, to generate finer initial change maps for RSIs. Shi et al. [38] proposed a deeply supervised attention metric-based network (DSAMNet), which designed a metric module to learn change maps by means of deep metric learning and the convolutional block attention module (CBAM) was integrated to provide more discriminative features. Lv et al. [39] introduced the spatial–spectral attention mechanism and multiscale dilation convolution strategy to reduce the pseudo changes and further enhance the detection performance. Wang et al. [40] designed a new adaptive spatial and channel fusion attention mechanism to enhance the changed features in the spatial and channel dimensions, which can completely obtain the relationships and differences between the features of bitemporal images. Jiang et al. [41] designed a co-attention module to emphasize the correlation among the input feature pairs, and integrated the attention module into the pyramid structure to obtain richer target information.

C. Transformer-Based Change Detection Method

Transformer was first applied in the field of natural language processing and has achieved good performance. Recently, transformer was introduced into the computer vision field. It is widely used in many tasks of the computer vision field, such as image segmentation [43], object detection [44], image generation [45], and image classification [46]. Unlike CNN which extracts information by constantly deepening the network and expanding the receptive field, the transformer interacts with the global information through the self-attention mechanism and can handle long-distance dependence problems. Hence, the transformer has a preliminary application in the task of RSI change detection.

Zhang et al. [23] designed a pure swin transformer network with a Siamese U-shaped structure, which was not limited by the intrinsic locality of convolution operation and can capture global information in the space-time dimension. Li et al. [25] introduced the transformer into the Unet structure, and encoded the tokenized image patches from the CNN feature map to model the context and obtain rich global context information. To tap the potential of integrating CNN and transformer, Feng et al. [27] shifted the design paradigm from series to parallel in order and proposed a dual-branch structure, in which CNN branch is used to extract local features and transformer is performed to capture the global dependencies. However, these transformer-based methods only consider the information in the space-time dimension and ignore the importance information in the spectral dimension. Therefore, these methods are not suitable for processing HSIs.

III. Method

The proposed SATNet consists of three modules: SETrans feature extraction module, the transformer-based correlation representation module and the detection module, as shown in Fig. 1. SETrans feature extraction module performs a series of squeeze units and expand units to extract deep features of HSIs. Transformer-based correlation representation module aims to explore the correlation of learned deep features between two HSIs from the perspective of difference and dot-product operations. Moreover, the transformer encoder in the transformer-based correlation representation module is employed to capture the long-distance dependence of spectral information. Finally, the detection module adopts the decision fusion strategy to weighted fuse the change detection map acquired from the differential stream, multiplicative stream, and their corresponding concatenating stream to generate the final change map.

A. SETrans Feature Extraction Module

The SETrans feature extraction module consists of two components: the squeeze and expand (SE) block and the transformer encoder.

TABLE I
STRUCTURE OF THE SE BLOCK

Unit	Input Size	Output Size
Squeeze Unit 1	7× 7× C	7× 7× 128
Squeeze Unit 2	7× 7× 128	7× 7× 64
Squeeze Unit 3	7× 7× 64	7× 7× 32
Squeeze Unit 4	7× 7× 32	7× 7× 16
Conv layer	7× 7× 16	7× 7× 16
Expand Unit 1	2× (7× 7× 16)	7× 7× 32
Expand Unit 2	2× (7× 7× 32)	7× 7× 64
Expand Unit 3	2× (7× 7× 64)	7× 7× 128
Expand Unit 4	2× (7× 7× 128)	7× 7× 256
Conv layer	7× 7× 256	7× 7× C

1) SE Block: The SE block is composed of a series of squeeze units and expand units, which is a simplified structure of dense connection [47]. In the process of feature extraction, it gradually reduces and restores the spectral features to remove noise and obtain representative features.

The structure of the SE block is shown in Table I. The SE block consists of four squeeze units, four expand units, and two Conv layers. The kernel size of both Conv2d layers is 3×3 with a stride of 1 and a padding size of 1.

The squeeze unit contains two sets of convolution layer (Conv2d): batch normalization layer (BN) and Relu activation layer (ReLu). Given an input image with the size of $7 \times 7 \times C_{\text{in}}$, the size of the output feature is $7 \times 7 \times C_{\text{out}}$, where 7×7 is the patch size. The kernel size of both Conv2d layers is 3×3 with a stride of 1 and a padding size of 1.

The expand unit is composed of a concatenating layer and two Conv2d+BN+ReLU layers. Given two input features of identical size ($7 \times 7 \times C_{\text{in}}$), the size of the input features in this unit is $2(7 \times 7 \times C_{\text{in}})$, and this unit outputs a feature map of size $7 \times 7 \times C_{\text{out}}$. The details of the Conv2d layers in the expand unit are identical to those in the squeeze unit. In addition, the expand unit includes a concatenation operation to recover the details of features.

It is noteworthy that the SE block includes only convolution, not pooling operations. Each pixel in HSIs can generally represent an area of several meters or even tens of meters on a side, which contains rich information. However, the pooling operation randomly discards pixels, which will lead to the loss of region information. Hence, we use convolutions to transform features in the spectral dimension, without changing the corresponding feature size under the spatial dimension. For the input bitemporal HSIs $\boldsymbol{T}_1$ and $\boldsymbol{T}_2 \in \mathbb{R}^{H\times W\times C}$, we can obtain the deep-level features F_{T1} and $F_{T2} \in \mathbb{R}^{H\times W\times C}$, respectively.

In addition, the transformer encoder is added to the SE block in the proposed SATNet to extract more representative features in the spectral dimension. The features obtained after the third and fourth squeeze units are more abstract and global.

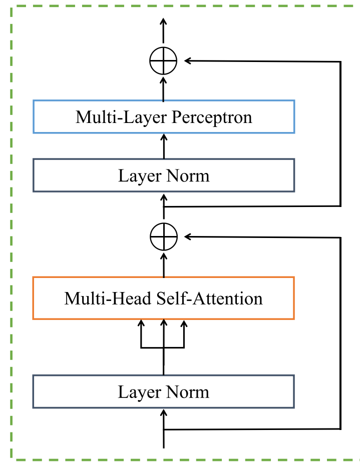


Fig. 2. Structure of the transformer encoder.

Therefore, the transformer encoder is only added to these two features to interact the spectral information with less noise, to better represent the regions of interest. The transformer encoder will be described in detail in the following.

2) Transformer Encoder: Transformer is based on self-attention mechanism, which interacts information through dot-product operation, and comprehensively considers all context information when processing a sequence. It can well address the problem of long-distance dependence because a self-attention operation can learn the global information obtained by multiple convolutions and poolings of CNN.

Transformer encoder is composed of L identical blocks connected in series. Each block mainly includes layernorm (LN), multi-head self-attention (MSA), multilayer perceptron (MLP), and residual connection, as shown in Fig. 2. For the input feature $\boldsymbol{F} \in \mathbb{R}^{B\times \text{HW}\times C}$($B$ is the batch size), each block will get an output $\boldsymbol{Z}_k \in \mathbb{R}^{B\times \text{HW}\times C}$ $(1 \leq k \leq L)$, and the final output of the transformer encoder $\boldsymbol{Y} \in \mathbb{R}^{B\times \text{HW}\times C}$ is the output of the last layer (i.e., $\boldsymbol{Z}_L$). The operations in a block can be expressed as

$$Z'_k = \text{MSA}(\text{LN}(Z_{k-1})) + Z_{k-1} \quad (1)$$
$$Z_k = \text{MLP}\big(\text{LN}\big(Z'_k\big)\big) + Z'_k. \quad (2)$$

The calculation formulas of self-attention are written as follows:

$$\text{Attention}(Q, K, V) = \sigma\left(\frac{\text{QK}^T}{\sqrt{d}}\right)V \quad (3)$$

where $Q, K, V \in \mathbb{R}^{C\times d}$ are the learnable parameters which can be obtained from the input, and d is the channel dimension. σ represents the activation function.

Then, MSA can be formulated as

$$\text{MSA}(Q, K, V) = \text{concat}(\text{head}_1, \ldots, \text{head}_h)W^O \quad (4)$$
$$\text{head}(i) = \text{Attention}\Big(\text{QW}_i^Q, \text{KW}_i^K, \text{VW}_i^V\Big) \quad (5)$$

where W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{C\times d}$, $W^O \in \mathbb{R}^{Hd\times C}$ are the linear projection matrices, and h is the number of heads.

It can be seen that the transformer encoder in the proposed SATNet is the same as the transformer encoder in vision transformer (ViT) [48]. But the transformer encoder in the proposed

SATNet focuses on mining the long-distance dependence in the spectral dimension of HSIs. HSIs have more than a dozen or even hundreds of spectral bands, containing rich spectral information, which provide the possibility for detecting changes from the spectral perspective. Since the transformer encoder used in the proposed SATNet aims to explore the long-distance dependence in the spectral dimension of HSIs, it does not need to do position encoding and generate the class token. However, since ViT deals with natural images that only contain red-green-blue (RGB) channels, the transformer encoder in ViT focuses on solving the long-distance dependence in the spatial dimension.

B. Transformer-Based Correlation Representation Module

The change detection of HSIs aims to explore the correlation of bitemporal HSIs or extracted corresponding features, and then generate change maps by measuring the similarity of extracted features of bitemporal HSIs. The difference and dot-product are two methods to calculate the correlation. The smaller the value achieved by the difference operation of two vectors, the two vectors are more similar. The larger the value obtained by dot-product of two vectors, the two vectors are more similar. However, the simple difference or dot-product operation is affected easily by noise. In this article, both difference and dot-product operation are used to construct the differential and multiplicative streams to explore the correlation of the learned features from the SETrans feature extraction module to alleviate the effect of noise. Moreover, the transformer encoder is adopted to explore the underlying information of the difference features and the dot-product representations to obtain more discriminative features.

In the transformer-based correlation representation module, we perform pixel-by-pixel subtraction and multiplication operations on the feature maps $\boldsymbol{F}_{T1}$ and $\boldsymbol{F}_{T2}$ extracted from the SETrans feature extraction module to obtain the difference map and the dot-product map. Then, the difference features and the dot-product features are fed into the transformer encoder to acquire difference features based on transformer encoder F'_{DS} and dot-product features based on transformer encoder F'_{MS}, which can mine the effective information in the spectral dimension to extract more discriminative features. The transformer encoder structure is the same as that in the SETrans feature extraction module.

The process of this module can be written as follows:

$$F'_{\mathrm{DS}} = \mathrm{TE}(\alpha_1 \times (F_{T1} \ominus F_{T2})) \tag{6}$$

$$F'_{\mathrm{MS}} = \mathrm{TE}(\alpha_2 \times (F_{T1} \otimes F_{T2})) \tag{7}$$

where TE represents the transformer encoder, α_1 and α_2 represent the weights of the two substreams, respectively. The output of this module F_{DS} and $F_{\mathrm{MS}} \in \mathbb{R}^{H\times W\times C}$ can be obtained from F'_{DS} and F'_{MS} after the reshape operation.

Because the results after subtraction and multiplication have significant numerical differences, so they need to be balanced by multiplying them with the appropriate weighting parameters α_1 and α_2. The features obtained after difference and dot-multiplied operations can represent the correlation, which is an important basis for the change detection task.

C. Detection Module

Two corresponding change features are generated by the differential and multiplicative streams, respectively. How to fuse these two change features is very important because it determines which information will be used more for the final change detection.

To make full use of the information learned from the differential and multiplicative streams and further explore the correlation of two streams, a decision fusion strategy is presented in our proposed SATNet to obtain the final change detection map, which focuses on fusing the change maps rather than deep features. The decision fusion strategy not only has stronger representation ability to obtain more useful information, but also can improve the robustness of the model.

We first perform a softmax function to generate the change maps of the differential stream and multiplicative stream, which are called $\mathrm{CM_{DS}}$ and $\mathrm{CM_{MS}} \in \mathbb{R}^{\mathrm{HW}\times 2}$, respectively. Then, the change map of the concatenating stream $\mathrm{CM_{con}} \in \mathbb{R}^{\mathrm{HW}\times 2}$ is obtained by a series of operations for learned features F_{DS} and F_{MS}, which can be written as

$$\mathrm{CM_{con}} = \sigma(\mathrm{FC}(\mathrm{concat}(F_{\mathrm{DS}}, F_{\mathrm{MS}}))) \tag{8}$$

where $\mathrm{concat}(a, b)$ represents the concatenating operation of the feature a and feature b, FC represents the fully connected network, and σ represents the softmax function.

The change maps detected from the differential stream, multiplicative stream, and concatenating stream are weighted and summarized to generate the final change map, which can be written as follows:

$$D_{\mathrm{total}} = w_1 \times \mathrm{CM_{DS}} + w_2 \times \mathrm{CM_{MS}} + w_3 \times \mathrm{CM_{con}} \tag{9}$$

where $w_i (1 \le i \le 3)$ are penalty parameters and represents the weights of the components and $\sum_{i=1}^{3} w_i = 1$.

D. Loss Function

The loss function used in this article is a compound loss function that mainly includes two terms: binary cross entropy loss (BCE) and contrastive loss (CL). The total loss can be written as

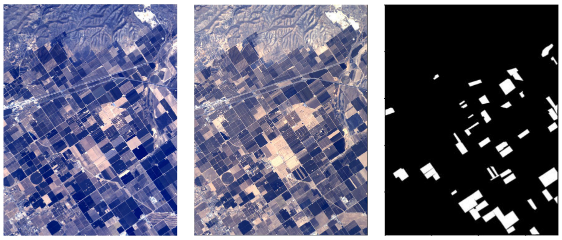
$$\mathrm{Loss} = \lambda_1 L_{\mathrm{BCE}} + \lambda_2 L_{\mathrm{CL}} \tag{10}$$

where λ_1 and λ_2 are penalty parameters that represent the weight of the BCE loss and CL, respectively.

BCE loss is the cross entropy loss for binary classification problems and is used when the label is zero or one. Unlike multiclass problems that require predicting a probability vector, only two classes of probabilities need to be predicted for binary classification problems. The BCE loss formula can be simplified

$$L_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right] \tag{11}$$

where the y_i and $\hat{y}_i (1 \le i \le N)$ represent the ground truth and the prediction, respectively. N is the total number of samples.



(a) (b) (c)

Fig. 3. Santa Barbara dataset. (a) HSI acquired in 2013. (b) HSI acquired in 2014. (c) Ground truth.

The mathematical expression of CL is written as follows:

$$L_{\mathrm{CL}} = \frac{1}{2N}\left[\mathrm{yd}^2 + (1-y)\max(\mathrm{margin} - D, 0)^2\right] \quad (12)$$

$$D = \|X_1 - X_2\|_2 \quad (13)$$

where y is the label, indicating whether two samples are similar. D is the Euclidean distance of features of two samples. The margin is the artificially set threshold, which is set to 1.5 in the experiment.

The aim of CL is to shorten the distance between similar samples and increase the distance between dissimilar samples. In the proposed SATNet, CL of $\boldsymbol{F}_{T1}$ and $\boldsymbol{F}_{T2}$ is calculated to measure the similarity of change maps. When two samples are unchanged, CL penalizes a large distance between them to ensure that similar samples are still similar after the feature extraction. When two samples are changed, CL punishes a small distance between them, as the larger the distance between them, the more dissimilar they are.

IV. Experiment

In this section, we will validate the effectiveness of the proposed SATNet on three real hyperspectral datasets. First, we will introduce three real datasets, the parameters setting in our experiments in detail and the evaluation criteria. Then, the comparison methods are introduced. The change detection performance of these comparison methods and our proposed SATNet are analyzed both quantitatively and qualitatively. Finally, we will conduct the ablation study to discuss the effectiveness of the two-stream structure and the transformer encoder.

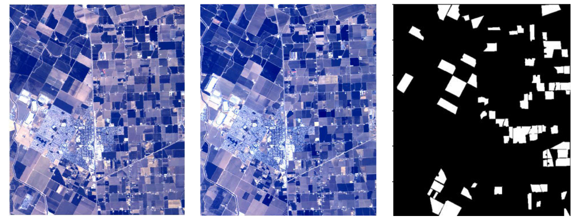
A. Datasets

Three bitemporal real hyperspectral datasets are Santa Barbara, Bay Area, and Hermiton City, which are acquired from airborne visible infrared imaging spectrometer (AVIRIS) sensor or HYPERION sensor.

Santa Barbara dataset was taken in Santa Barbara area (California) in 2013 and 2014 via AVIRIS sensors. The size of HSI is 984×740 pixels with 224 bands. Two images and the ground truth are shown in Fig. 3.

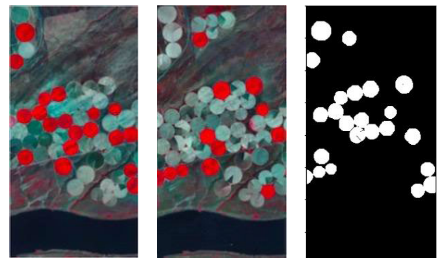
Bay Area dataset was acquired in Paterson area (California) in 2013 and 2015 by AVIRIS sensors. The size of HSI is 600×500 pixels with 224 bands. Two images and the ground truth are shown in Fig. 4.

Hermiston City dataset was taken in Hermiston area (Oregon) in 2004 and 2007 through HYPERION sensors. The size of HSI is 390×200 pixels with 242 bands. Two images and the ground truth are shown in Fig. 5.



(a) (b) (c)

Fig. 4. Bay Area dataset. (a) HSI acquired in 2013. (b) HSI acquired in 2015. (c) Ground truth.



(a) (b) (c)

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

In the pre-processing stage, we directly split the image into $7 \times 7 \times C$ patches pixel by pixel. The specific steps for the dataset generation are as follows: first, the specular reflection is performed on the original image to generate patches for edge pixels, and then a sliding window operation with a stride size of one is used to obtain patches for all pixels. Specifically, if the size of HSI is $H \times W \times C$, we can obtain $H \times W$ patches of size $7 \times 7 \times C$. The reason for splitting HSI into patches of size 7×7 is that each pixel of HSI contains rich spectral information, and a patch with a larger patch size tends to contain more land covers because of low spatial resolution characteristic of HSIs.

Since the number of unchanged samples is much larger than that of changed samples in the dataset, the changed and unchanged samples are selected at the ratio of 1:1 to construct a sub-dataset to better reduce the impact of class imbalance. Eighty percentage of samples in the constructed sub-dataset are selected to build the training set. The remaining samples in the whole dataset were used as the testing dataset, and the whole dataset was utilized to calculate evaluation metrics and display qualitative results.

B. Experiment Setup

1) Parameter Setting: The proposed SATNet is implemented by the Pytorch framework, using a single NVIDIA GeForce GTX 3080Ti GPU with 12G memory for training and testing. In this article, Adam is selected as the optimizer to optimize the loss. The learning rate is 1E-5. The batch size is 32, and the training epoch is 50.

The weighted parameters α_1 and α_2 in the transformer-based correlation representation module are designed to balance the contributions of features acquired from difference and dot-multiplied operations to the change detection performance.

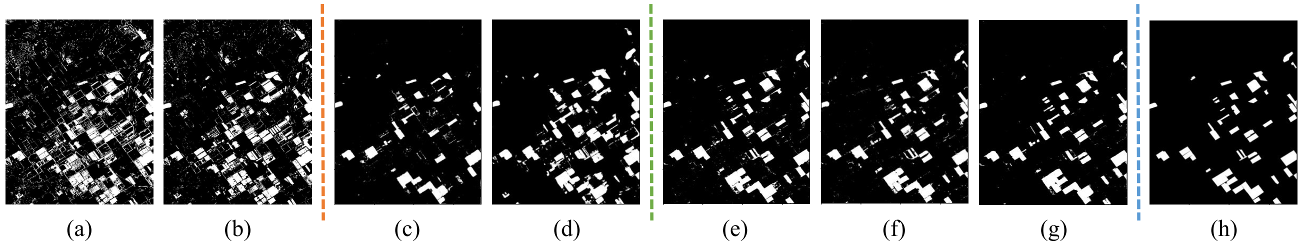


Fig. 6. Experiment result on Santa Barbara dataset. (a) CVA. (b) RCVA. (c) DSAMNet. (d) SSAN. (e) PA-Former. (f) BIT. (g) SATNet(ours). (h) Ground truth.

Since it was found experimentally that the feature values obtained by the dot-multiplied operation are about ten times higher than that obtained by the difference operation, so we set $\alpha_1 = 10$, $\alpha_2 = 1$ to equalize their contributions.

In the detection module, since the concatenating stream contains information from both the differential stream and multiplicative stream, we set $w_3 \geq 0.5$. Moreover, the difference operation is a commonly used method to measure similarity in the change detection task, so we set $w_1 \geq w_2$. The best performance is achieved based on experimental results when $w_1 = 0.4$, $w_2 = 0.1$, and $w_3 = 0.5$.

Since the change detection task is essentially a binary classification task, the penalty parameter of BCE loss λ_1 is set to one. The penalty parameter of CL λ_2 was adjusted to seek the best combination of penalty parameters λ_1 and λ_2. The best change detection performance is obtained based on experimental results when $\lambda_1 = 1$ and $\lambda_2 = 0.25$.

2) Evaluation Criteria: There are four evaluation indicators used in this article: OA, kappa coefficient, precision, and F1.

OA is the overall accuracy and the proportion of the correctly predicted pixels to the number of pixels in the whole image. The formula is written as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{14}$$

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

Kappa coefficient is an accuracy measure of the classification task. The calculation of the coefficient is based on the confusion matrix, and the value is between -1 and 1. The closer the result is to one, the better the consistency of the classification. The formula can be written as follows:

$$\mathrm{Kappa} = \frac{\mathrm{OA} - \mathrm{P_e}}{1 - \mathrm{P_e}} \tag{15}$$

$$\mathrm{P_e} = \frac{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN}) + (\mathrm{FN} + \mathrm{TN})(\mathrm{FP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})^2}. \tag{16}$$

Precision and Recall can well reveal the problem of false detection and miss detection of algorithms. The calculation formulas of the precision and recall are written as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{17}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{18}$$

F1 score is the harmonic mean of the precision and recall, so this criterion can balance the impact of precision and recall, and reflect the average performance of the model. The formula can be written as follows:

$$\mathrm{F1} = \frac{2}{\frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}}} = \frac{2 \times (\mathrm{Precision} \times \mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}}. \tag{19}$$

C. Comparison Results and Analysis

1) Competitors:

1) CVA [49]: The CVA method generates a new difference hyperspectral image by calculating the difference between bitemporal images, and then sets a threshold value for the intensity of changes according to specifics of the region.
2) RCVA [50]: Robust CVA is based on the CVA method, which considers neighbors of the target pixel to enhance the robustness to differences in viewing geometries or noise from registration.
3) DSAMNet [51]: DSAMNet performs the CBAM module to learn more discriminative features from the backbone, and generates the change map through a change decision module using the metric learning strategy.
4) Spectral and spatial attention network (SSAN) [52]: SSAN extracts spectral-spatial features through a joint attention network with a recurrent neural network (RNN) attention branch in the spectral dimension and a CNN attention branch int the spatial dimension. Then, a fusion module is performed to obtain the final change detection results.
5) Prior-aware Former (PA-Former) [24]: PA-Former designs a prior-feature extractor to obtain prior and deep features from bitemporal images, and a prior-aware transformer module is performed to capture cross-temporal and long-range contextual information.
6) Bitemporal image transformer (BIT) [22]: BIT uses visual words (semantic tokens) to represent high-level concepts of changes of interest and leverages the transformer structure to efficiently and effectively model contexts within the spatial and temporal domains.

2) Experiment on Santa Barbara: The detection results of the Santa Barbara dataset are visualized in Fig. 6. It can be seen that traditional change detection methods (i.e., CVA and RCVA) have a lot of noise, especially at the top of the change map. Compared with traditional methods, the top of change maps acquired from two attention-based methods DSAMNet and SSAN have few noises. However, in the middle and lower parts of change maps, DSAMNet, and SSAN have a serious problem of miss detection and false detection, respectively.

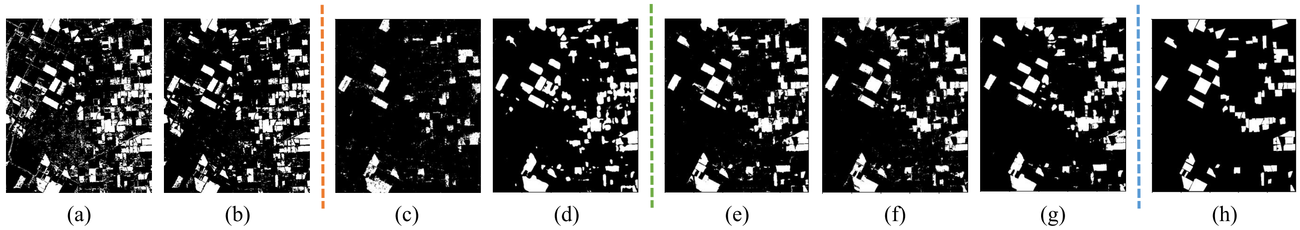


Fig. 7. Experiment result on Bay Area dataset. (a) CVA. (b) RCVA. (c) DSAMNet. (d) SSAN. (e) PA-Former. (f) BIT. (g) SATNet(ours). (h) Ground truth.

TABLE II
QUANTITATIVE RESULTS OF THE FOUR CRITERIA ON THE SANTA BARBARA DATASET

Method	OA	kappa	precision	F1
CVA	0.8677	0.4216	0.3345	0.4812
RCVA	0.8987	0.4926	0.4007	0.5417
DSAMNet	0.9602	0.6797	0.7593	0.7008
SSAN	0.9462	0.6851	0.5769	0.7129
PA-Former	0.9814	0.8746	0.7970	0.8845
BIT	0.9802	0.8657	0.7934	0.8762
SATNet(ours)	**0.9839**	**0.8883**	**0.8302**	**0.8969**

TABLE III
QUANTITATIVE RESULTS OF THE FOUR CRITERIA ON THE BAY AREA DATASET

Method	OA	kappa	precision	F1
CVA	0.8544	0.5091	0.4673	0.5901
RCVA	0.8884	0.5835	0.5520	0.6474
DSAMNet	0.9205	0.5881	0.8053	0.6303
SSAN	0.9297	0.7062	0.7066	0.7469
PA-Former	0.9563	0.8231	0.7781	0.8483
BIT	0.9565	0.8176	0.7781	0.8427
SATNet(ours)	**0.9630**	**0.8448**	**0.8226**	**0.8662**

The transformer-based methods PA-Former, BIT, and the proposed SATNet have lower false and missed detection rates compared with the attention-based methods. The proposed SATNet achieves the best detection result because it uses the transformer mechanism in the spectral dimension to explore the spectral dependency of spectrum.

The quantitative evaluation of the performance of the proposed SATNet on Santa Barbara dataset is shown in Table II. The performance of traditional methods in hyperspectral change detection task is not good enough, the indicators kappa, precision and F1 score are far lower than other methods. The performance of deep learning-based methods is much better than that of traditional methods. One important reason is that they all use neural networks to extract deep features that are representative. Two attention-based methods DSAMNet and SSAN have better results on OA, kappa, and F1 score, because they use the attention mechanism to focus on the change areas. But the precision of SSAN is only 0.5769, which is the lowest among deep learning methods, indicating that this method has the high false detection rate. SATNet has achieved the best results in all indicators, reaching 0.9839, 0.8883, 0.8302, and 0.8969, respectively. This is because SATNet performs the decision fusion strategy to make full use of learned features, which can obtain more accurate change maps and reduce the impact of noises.

3) Experiment on Bay Area: The final change detection results of the competitors and the proposed method on the Bay Area dataset are shown in Fig. 7. CVA and RCVA have poor performance due to their manual features. DSAMNet and SSAN generate the weight mask of features by the spatial attention or spectral attention module to enhance the importance of specific regions to obtain more representative features. Transformer conducts global interaction through the self-attention mechanism, so that each patch contains global features, which further improves the detection results. However, the dataset contains many small change areas and these change regions are scattered, PA-Former and BIT did not achieve good results on this dataset. SATNet performs the transformer-based correlation representation module to capture the correlation of learned deep features from the perspective of difference and dot-product operations, obtaining more discriminative features and achieving the best detection results among these methods.

Table III shows quantitative results on the Bay Area dataset. SATNet performs best in four indicators, and the kappa, precision, and F1 score are more than 0.82. The precision of CVA and RCVA only reach 0.4673 and 0.5520, respectively. Although attention-based methods have greatly improved in OA (greater than 0.92), they still perform poorly in Kappa and F1 score. PA-Former and BIT are significantly lower than SATNet in three indicators, with an average decrease of 0.03, indicating that the prediction results of these two models are not accurate enough. It is because they ignore the importance of spectral information and lack the interaction between spectral information.

4) Experiment on Hermiston City: Fig. 8 depicts the change detection results for the Hermiston City dataset. The traditional methods CVA and RCVA have a relatively high false detection rate and perform poorly compared with deep learning-based methods. The attention-based method DSAMNet has a high false detection rate, while SSAN has a high miss detection rate. This indicates that the attention mechanism does not perform well in this dataset and is inadequate for hyperspectral image feature extraction. The results of PA-Former, BIT, and SATNet based on transformers are good. However, there is a small amount of noise in the PA-Former result, and the edge part of the BIT result is not smooth enough. SATNet can well alleviate these two problems, because PA-Former and BIT extract features in the spatial dimension, while SATNet interacts information in the spectral dimension. Since the spectral

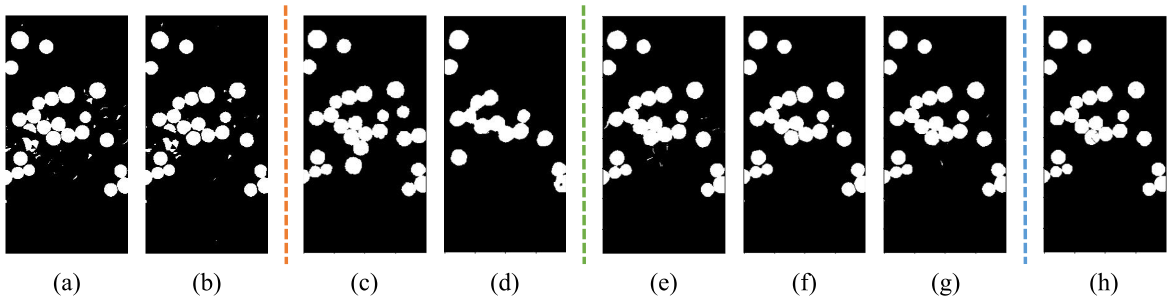


Fig. 8. Experiment result on Hermiston City dataset. (a) CVA. (b) RCVA. (c) DSAMNet. (d) SSAN. (e) PA-Former. (f) BIT. (g) SATNet(ours). (h) Ground truth.

TABLE IV
QUANTITATIVE RESULTS OF THE FOUR CRITERIA ON THE HERMISTON CITY DATASET

Method	OA	kappa	precision	F1
CVA	0.9241	0.5959	0.4828	0.6336
RCVA	0.925	0.5893	0.4857	0.6271
DSAMNet	0.9680	0.8697	0.8038	0.8881
SSAN	0.9701	0.8102	0.706	0.8259
PA-Former	0.9881	0.9487	0.9195	0.9555
BIT	0.9895	0.9541	0.9343	0.9601
SATNet(ours)	**0.9929**	**0.9690**	**0.9556**	**0.9731**

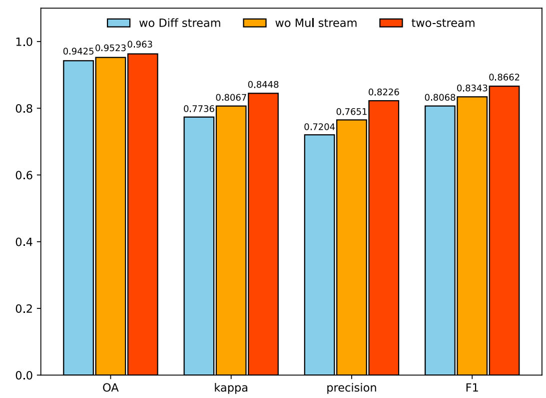


Fig. 9. Ablation results of different streams used in the transformer-based correlation representation module.

information of hyperspectral images in spectral dimension is rich and very important for the change detection task, interacting spectral information can achieve better results.

The quantitative results for Hermiston City dataset are shown in Table IV. OA, kappa, precision, and F1 score values of SATNet are 0.9929, 0.9690, 0.9556, and 0.9731, respectively. The quantitative performance of traditional methods is unsatisfactory, and the quantitative results of DSAMNet and SSAN via the attention mechanism are improved but still limited. Compared with the other two datasets, Hermiston city dataset is simpler, and three transformer-based methods have achieved better results. SATNet has achieved the best result due to its sensitivity to spectral information, which is obtained through the global interaction and the long-range dependencies of spectral information.

D. Ablation Study

1) Effective of Two-Stream Structure: To verify the effectiveness of the two-stream structure in the transformer-based correlation representation module, the following ablation experiments are designed. We remove the differential stream or multiplicative stream, denoted as “wo Diff stream” and “wo Mul stream,” respectively, leaving only one branch to generate the final change map. The ablation experiments are carried out on the Bay Area datasets, and the results are shown in Fig. 9.

It can be seen that after removing the differential stream or multiplicative stream respectively, the performance of the model decreases by 0.0712 and 0.0381 on kappa, and 0.1022 and 0.0575 on precision. It fully illustrates the effectiveness of the two-stream structure, which can better mine the correlation and relationship of learned deep features from difference and dot-product operations, and increase the representation ability of the proposed SATNet.

2) Effective of Transformer Encoder: To verify the effectiveness of transformer encoder, we design other two structures to process features after subtraction and multiplication, as shown in Fig. 10. Fig. 10(a) shows that no operation is performed after the two-stream structure, and features are directly fed to the detection module for the change detection. Fig. 10(b) represents that a CNN structure is designed to replace the transformer encoder, the structure contains two CNN blocks. Each block contains two convolutional layers and a maximum pooling layer, and then features extracted by CNN are fed to the detection module. Fig. 10(c) is the structure used in the proposed SATNet.

The ablation experimental results on the Bay Area dataset are shown in Fig. 11. The results of structure (a) on kappa and precision only reach 0.4212 and 0.4478, indicating that if features are not further processed after the two-stream structure, the final detection results are not good. If CNN structure is used to extract features, the performance can be greatly improved, but it is still lower than that of the transformer encoder. The kappa and precision values of structure (b) are 0.8076 and 0.7371, which are lower than that of structure (c), respectively. It is because CNN only extracts features in the spatial dimension, without considering the correlation of learned features in the spectral dimension. The transformer encoder can fully interact spectral information and improve the discriminability of features.

In addition, we also conducted corresponding ablation experiments on the transformer encoder in the SE block to

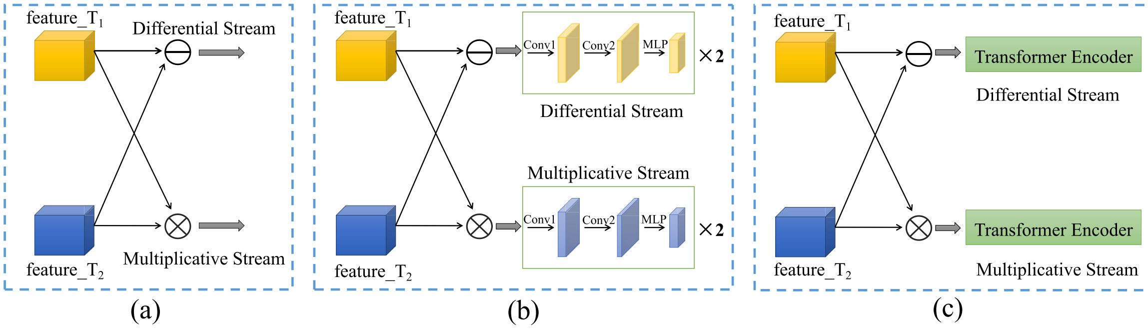


Fig. 10. Three structures to process features after subtraction and multiplication. (a) No operation is performed after the two-stream structure. (b) CNN structure is designed to replace the transformer encoder. (c) Structure used in the proposed SATNet.

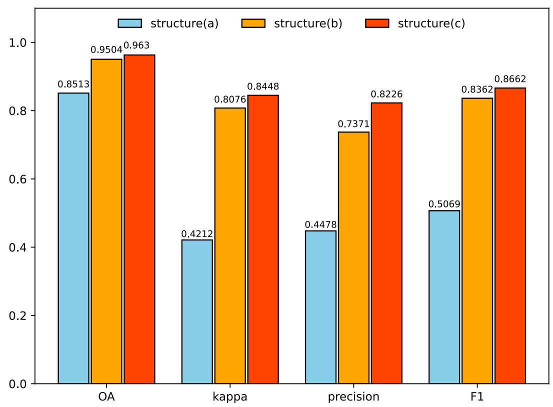


Fig. 11. Ablation results of three structures used in the network.

TABLE V

EXPERIMENT RESULTS OF THE DIFFERENT STRUCTURE ON THE BAY AREA DATASET

structure	OA	kappa	precision	F1
without TE	0.9459	0.7878	0.7292	0.8190
TE in all layers	0.9556	0.8167	0.7878	0.8424
TE in $3^{th}/4^{th}$ layers	0.9630	0.8448	0.8226	0.8662

explore that adding transformer encoder to which layer can make the model achieve the best performance. Hence, we use two other structures without the transformer encoder in the SE block and with the transformer encoder after all the squeeze blocks to accomplish the ablation study. The results on Bay Area dataset can be shown in Table V.

It can be seen from Table V that when transformer encoder is not added to the SE block, OA, kappa, precision, and F1 score values decreased by 0.0171, 0.057, 0.0934, and 0.0472, respectively, indicating that adding transformer encoder to the SE block can improve the model performance. However, when the transformer encoder is added after all the squeeze blocks, the performance of the model decreases slightly. It means that if spectral dimension information has interacted in all squeeze processes, it may lead to insufficient extraction of features in the spatial dimension, and ultimately reduce the detection performance. When the transformer encoder is added to the third and fourth layers, the high-level features already have large receptive fields and contain more spatial information. Then, better results can be obtained by mining the spectral information.

TABLE VI

ABLATION RESULTS WITH OR WITHOUT SETRANS FEATURE EXTRACTION MODULE

Dateset	Structure	OA	kappa	precision	F1
Santa Barbara dataset	with SETrans	0.9839	0.8302	0.7292	0.8969
	wo SETrans	0.9453	0.6825	0.5723	0.7108
Bay Area dataset	with SETrans	0.9630	0.8448	0.8226	0.8662
	wo SETrans	0.9365	0.5110	0.5597	0.5451
Herminston City dataset	with SETrans	0.9929	0.9690	0.9556	0.9731
	wo SETrans	0.9211	0.8758	0.5723	0.9956

3) Effective of SETrans Feature Extraction Module: To verify the effectiveness of the SETrans feature extraction module, two ablation studies are carried out. First, the entire SETrans feature extraction module is removed, and the change detection results on three datasets are shown in Table VI. In Table VI, "with SETrans" means that the structure of SATNet incorporates the SETrans feature extraction module, while "wo SETrans" denotes the absence of the SETrans feature extraction module in the structure of SATNet. The term "wo SETrans" also implies that the differences and dot-products of image pairs are directly fed into the transformer encoder. Second, some layers within the SE block are removed, and the change detection results of the different numbers of SE units in the SE block for the Bay area dataset are shown in Fig. 12.

It can be seen from Table VI that the SETrans feature extraction module positively influences the overall performance of our proposed method. The removal of this module led to a decline in performance with respect to OA, kappa, precision, and F1 values. This indicates that the SE operation is crucial for the proposed method as it effectively captures the interdependencies among channels. Furthermore, Table VI reveals that while the transformer encoder can act as a feature extraction module, its performance is inferior compared to the original SATNet design. This suggests that the features extracted by the SETrans feature extraction module, which are learned via CNNs, provide more useful information for the change detection task.

It can be seen from Fig. 12 that all detection indicators, including OA, kappa, precision, and F1, have declined, with the reduction in the number of SE units in the SE block. This indicates that the proposed SATNet requires a sufficient number of SE units to achieve the optimal performance. When the number of SE units is four, the proposed network achieves

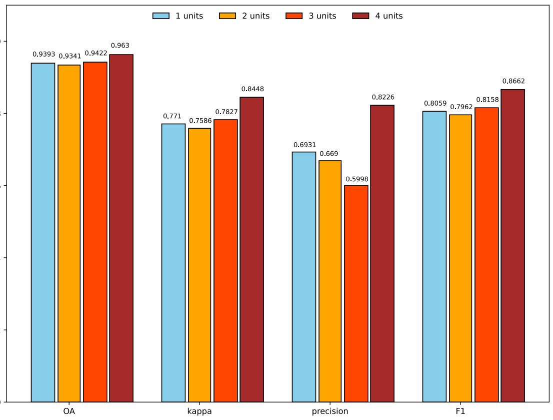


Fig. 12. Experiment results of the different numbers of SE units for the Bay Area dataset.

the best change detection results. Therefore, the number of SE units is set to four in our experiment.

V. Conclusion

In this article, we propose a SATNet for HSI change detection, which can explore the spectral dependence of the spectrum to extract more discriminative features. SATNet includes three components: SETrans feature extraction module, transformer-based correlation representation module, and detection module. The SETrans feature extraction module uses the SE block as the backbone and adds the transformer encoder after the squeeze layer to extract deep features. The transformer-based correlation representation module performs pixel-by-pixel subtraction and dot-multiplied operations to learn deep features, and then the differential and dot-product features are fed to the transformer encoder to extract more discriminative features in the spectral dimension. The detection module employs a new decision fusion strategy to obtain a more accurate final change detection map by weighted summing the detection results of differential stream, multiplicative stream, and their corresponding concatenating stream. The effectiveness of our proposed SATNet was validated on three real hyperspectral datasets. The experimental results show that the proposed DATNet achieves the best detection results compared with other existing methods.

References

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

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

[3] 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.

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

[5] P. de Bem, O. de Carvalho Junior, R. F. Guimarães, and R. T. Gomes, "Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks," *Remote Sens.*, vol. 12, no. 6, p. 901, Mar. 2020.

[6] K. Isaienkov, M. Yushchuk, V. Khramtsov, and O. Seliverstov, "Deep learning for regular change detection in Ukrainian forest ecosystem with Sentinel-2," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 364–376, 2021.

[7] L. Ke, Y. Lin, Z. Zeng, L. Zhang, and L. Meng, "Adaptive change detection with significance test," *IEEE Access*, vol. 6, pp. 27442–27450, 2018.

[8] B. Cui, Y. Zhang, L. Yan, J. Wei, and Q. Huang, "A SAR change detection method based on the consistency of single-pixel difference and neighbourhood difference," *Remote Sens. Lett.*, vol. 10, no. 5, pp. 488–495, May 2019.

[9] S. Ertürk, "Fuzzy fusion of change vector analysis and spectral angle mapper for hyperspectral change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 5045–5048.

[10] S. Achour, M. C. Elmezouar, N. Taleb, K. Kpalma, and J. Ronsin, "A PCA-PD fusion method for change detection in remote sensing multi temporal images," *Geocarto Int.*, vol. 37, no. 1, pp. 196–213, 2022.

[11] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J.-Y. Tourneret, "A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 799–812, Mar. 2015.

[12] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Sci. Inform.*, vol. 12, no. 2, pp. 143–160, Jun. 2019.

[13] H. Li, M. Gong, Q. Wang, J. Liu, and L. Su, "A multiobjective fuzzy clustering method for change detection in SAR images," *Appl. Soft Comput.*, vol. 46, pp. 767–777, Sep. 2016.

[14] T. Bao, C. Fu, T. Fang, and H. Huo, "PPCNET: A combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1797–1801, Oct. 2020.

[15] J. Zheng et al., "MDESNet: Multitask difference-enhanced Siamese network for building change detection in high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 15, p. 3775, Aug. 2022.

[16] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

[17] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[18] Z. Zheng, Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang, "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 247–267, May 2021.

[19] R. Liu, D. Jiang, L. Zhang, and Z. Zhang, "Deep depthwise separable convolutional network for change detection in optical aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1109–1118, 2020.

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

[21] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

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

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

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

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

[26] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 207–210.

[27] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410213.

[28] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[29] 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.

[30] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[31] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[32] W. Wiratama, J. Lee, S.-E. Park, and D. Sim, "Dual-dense convolution network for change detection of high-resolution panchromatic imagery," *Appl. Sci.*, vol. 8, no. 10, p. 1785, Oct. 2018.

[33] C. Yu, S. Zhou, M. Song, and C.-I. Chang, "Semisupervised hyperspectral band selection based on dual-constrained low-rank representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[34] C. Yu, C. Liu, H. Yu, M. Song, and C.-I. Chang, "Unsupervised domain adaptation with dense-based compaction for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12287–12299, 2021.

[35] V. Sadeghi, F. F. Ahmadi, and H. Ebadi, "A new fuzzy measurement approach for automatic change detection using remotely sensed images," *Measurement*, vol. 127, pp. 1–14, Oct. 2018.

[36] 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.

[37] J. Zhao, Y. Chang, J. Yang, Y. Niu, Z. Lu, and P. Li, "A novel change detection method based on statistical distribution characteristics using multi-temporal PolSAR data," *Sensors*, vol. 20, no. 5, p. 1508, Mar. 2020.

[38] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.

[39] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412712.

[40] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, Sep. 2021, Art. no. 102348.

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

[42] K. Song and J. Jiang, "AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4816–4831, 2021.

[43] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.

[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 213–229.

[45] B. Zhang et al., "StyleSwin: Transformer-based GAN for high-resolution image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11294–11304.

[46] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.

[47] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

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

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

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

[51] M. Liu and Q. Shi, "DSAMNet: A deeply supervised attention metric based network for change detection of high-resolution images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 6159–6162.

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

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

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

Liangxu Su received the bachelor's degree in network engineering from the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, China, in 2022. He is currently pursing the master's degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include deep learning, computer vision, remote sensing, and person search.

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

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

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