An Unsupervised Transformer-Based Multivariate Alteration Detection Approach for Change Detection in VHR Remote Sensing Images

Yizhang Lin, Sicong Liu, *Senior Member, IEEE*, Yongjie Zheng, *Student Member, IEEE*, Xiaohua Tong, *Senior Member, IEEE*, Huan Xie, *Senior Member, IEEE*, Hongming Zhu, *Member, IEEE*, Kecheng Du, Hui Zhao, and Jie Zhang

***Abstract*—Multitemporal change detection (CD) plays a crucial role in the remote sensing application field. In recent years, supervised deep learning methods have shown excellent performance in detecting changes in very-high-resolution (VHR) images. However, these methods require a large number of labeled samples for training, making the process time-consuming and labor-intensive. Unsupervised approaches are more attractive in practical applications since they can produce a CD map without relying on any ground reference or prior knowledge. In this article, we propose a novel unsupervised CD approach, named transformer-based multivariate alteration detection (trans-MAD). It utilizes a pre-detection strategy that combines the compressed change vector analysis and the iteratively reweighted multivariate alteration detection (IR-MAD) to generate reliable pseudotraining samples. More accurate and robust CD results can be achieved by leveraging the IR-MAD to detect insignificant changes and by incorporating the transformer-based attention mechanism to model the difference or similarity between two distant pixels in an image. The proposed trans-MAD approach was validated on two VHR bitemporal satellite remote sensing datasets, and the obtained experimental results demonstrated its superiority comparing with the state-of-the-art unsupervised CD methods.**

***Index Terms*—Change detection (CD), deep learning, iteratively reweighted multivariate alteration detection (IR-MAD), transformer, unsupervised, very-high-resolution (VHR) remote sensing images.**

Manuscript received 31 October 2023; revised 12 December 2023; accepted 31 December 2023. Date of publication 4 January 2024; date of current version 19 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42071324 and Grant 42241130, in part by the Shanghai Rising-Star Program under Grant 21QA1409100, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, in part by the Fundamental Research Funds for the Central Universities, and in part by the Research Project of Tongji Architectural Design (Group) Company Ltd. under Grant 2023J-JB11. *(Corresponding author: Sicong Liu.)*

Yizhang Lin, Sicong Liu, Xiaohua Tong, Huan Xie, Kecheng Du, Hui Zhao, and Jie Zhang are with the College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China (e-mail: evan@tongji.edu.cn; sicong.liu@tongji.edu.cn; xhtong@tongji.edu.cn; huanxie@tongji.edu.cn; kecheng_du@tongji.edu.cn; zhaohui@tongji.edu.cn; zhangjie22@tongji.edu.cn).

Yongjie Zheng was with the College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China. She is now with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: yongjie.zheng@unitn.it).

Hongming Zhu is with the School of Software Engineering, Tongji University, Shanghai 201804, China (e-mail: zhu_hongming@tongji.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3349775

I. Introduction

CHANGE detection (CD) is a crucial task in various remote sensing application fields. It involves identifying the differences of an object or phenomenon over a certain region by analyzing two or more images captured at different times [1]. CD technique enables successful applications like land-cover mapping, disaster assessment, urban development monitoring, and ecological environment monitoring [2]. It provides a great opportunity to discover and analyze land-cover changes caused by human activities or natural phenomenon, which helps to make prompt and sound decisions [3]. With the rapid development of observation platforms and optical sensors, the very-high-resolution (VHR) remote sensing images have become the primary data source in the data archive. VHR images contain fine spatial information and thus are able to depict land objects at a detailed scale, whereas their spectral information is relatively coarse when comparing with the dense-sampling hyperspectral images.

In the past decades, numerous CD methods have been proposed, ranging from traditional techniques to advanced deep learning-based approaches. The category of traditional CD methods includes image algebra, image transformation, post-classification comparison, and others [4]. Within this context, popular CD methods are mainly developed based on spectral analysis of the original bands, such as change vector analysis (CVA) [5], compressed change vector analysis (C^2VA) [6] and its adaptively sequential version (S^2CVA) [7], multivariate alteration detection (MAD) [8] and its iteratively reweighted version (IR-MAD) [9], principal component analysis (PCA) based [10], and slow feature analysis (SFA) based [11]. In order to explore more robust change representation, some works also focused on features that derived from the original spectral bands and incorporated into C^2VA or IR-MAD [12], [13]. However, such features are mainly shallow and artificially designed, whose effectiveness is not sufficient for representing different types of changes at different significance levels. Furthermore, these traditional CD methods may face challenges concerning reduced accuracy and robustness when dealing with the VHR images over complex scenes, since they are mainly designed based on the utilization of original spectral bands or extraction of simple handcrafted features. In addition, data quality issues, such as

the seasonable variations, illumination conditions, and spectral variability may lead to the degradation of the CD performance [14], [15], [16].

In recent years, deep learning techniques have demonstrated remarkable success in various remote sensing application tasks [17], [18], [19], [20], [21], [22], [23]. They have also emerged as promising alternatives to address the limitations of traditional CD methods. According to the utilization of reference data, they can be divided into two main categories: supervised and unsupervised methods. The former relies on the available ground reference samples to train the model, such as fully convolutional network (FCN) [24], UNet++ [25], and bitemporal image transformer (BIT) [26]. The latter does not require ground reference samples thus is data-driven with a higher degree of automation. Gong et al. [27] proposed a method based on a generative adversarial network (GAN) to generate better differential images, and built the mapping relationship between training data and corresponding patches, and finally obtained binary change maps. Saha et al. [28] developed a deep change vector analysis (DCVA) framework, and made full use of the multi-layer deep features extracted by convolutional neural network (CNN) to determine the changed pixels. Chen et al. [29] proposed a method named DSMS-CN that used the deep siamese CNN to extract the multi-scale spectral-spatial features for CD. Du et al. [30] proposed a deep SFA (DSFA) method based on the original SFA, which used two symmetric fully connected networks to project input data into a new feature space, and then extracted the most invariant components to highlight the changed components. Wu et al. [31] built a deep siamese kernel principal component analysis convolution mapping network (KPCA-MNet) to extract high-level spectral-spatial feature maps and generated the final CD map.

Despite the effectiveness of the aforementioned unsupervised approaches, several issues still require to be analyzed and addressed.

First, the quality of pseudotraining samples largely depends on the pre-detection step. If the pre-detection algorithm does not work properly, it will introduce inaccurate or even wrong samples during the training process, eventually affecting CD accuracies. For example, in the existing KPCA-MNet algorithm, image patches are randomly selected as samples for training the KPCA convolutional layer, whose uncertainty will inevitably lead to unstable CD performance with the occurrence of omission and commission errors. The predetection step in DSFA is achieved by using CVA to obtain pseudotraining samples. To follow the convention of SFA, it only selects the pixels with the lowest change intensity as samples. However, this cannot be successfully extended to other methods relying on different CD strategies. Note that the insufficient representativeness of the samples will result in poor performance of CD especially in those complex scenarios.

Second, existing methods exhibit limitations in feature extraction. DCVA relies on pretrained deep CNN to extract deep features, and its generalization ability is unstable when dealing with different complex scenarios. The DSMS-CN method utilizes deep siamese CNNs to extract multiscale spectral-spatial features. However, due to its inherent characteristics and limited receptive field of convolutional kernels, CNN can only capture spatial contextual information at the local scale. It is challenging to capture long-range dependencies and contextual information, which has a significant impact on CD accuracy, leading to the occurrence of commissions and omissions. There are existing methods (e.g., BIT, hybrid-TransCD [32], and TransUNetCD [33]) using Transformer to model global contextual features, but they are all trained in a supervised manner and require a large number of training samples.

To overcome these limitations, in this article, we propose a novel deep learning-based CD framework named transformer-based multivariate alteration detection (trans-MAD). The main contributions and novelties are summarized as follows.

1) *Improved Predetection* (IPD) *and Pseudotraining Sample Generation:* The challenging issue of pseudotraining samples generation in the unsupervised CD is addressed by taking advantages of the joint change representation from two independent pre-detection algorithms C^2VA and IR-MAD, without relying on the ground reference data or prior knowledge. This facilitates the generation of highly reliable and representative pseudotraining samples in an automated and unsupervised fashion, even in complex scenarios.
2) *Innovative Shallow-to-Deep* (SD) *Feature Extraction:* Considering that the traditional IR-MAD method directly performs linear transformations on the original spectral bands to extract change information, it is easily affected by noise. This article develops an innovative unsupervised framework that uses CNN, transformer, and IR-MAD to harmoniously extract and fuse SD features, effectively describing local and global change information, thereby reducing omission errors and generating more accurate CD maps.
3) *Robust Handling of Sampling Randomness:* Trans-MAD incorporates a dedicated decision fusion (DF) based CD module to mitigate the challenges posed by random sampling uncertainty when generating pseudotraining samples. This strategy reduces commissions in CD output and ensures the stability of the final result.

The proposed trans-MAD approach is validated on two real VHR image datasets, and obtained experimental results confirm its effectiveness when comparing with the several state-of-the-art unsupervised CD methods.

The rest of this article is organized as follows. Section II introduces the related works. Section III presents in detail the proposed approach. Section IV describes the datasets and detailed experimental design. Experimental results and discussions are provided in Section V. Finally, Section VI concludes this article.

II. Related Works

A. Convolutional Neural Network

The CNN, a specialized neural network architecture, is particularly designed for data with grid-like structures. It has emerged as a powerful tool for detecting changes in VHR remote sensing images. Many research endeavors involve training CNN models

to learn feature difference between bitemporal images, thereby facilitating the CD process [10].

CNN is characterized by hierarchical structure, typically comprising convolutional layers, activation functions, and pooling layers. They operate on input data using learnable convolution kernels, and apply activation functions to introduce nonlinearity to the network. Subsequently, the pooling layer subsamples each feature map to reduce redundancy. Over a series of alternating convolutional and pooling layers, CNN autonomously generates advanced features from the input data. Commonly, the network parameters are optimized using stochastic gradient descent (SGD) and the backpropagation algorithm. Through several rounds of training with annotated data, CNN performs supervised learning and generates increasingly representative features.

B. Bitemporal Image Transformer

The BIT serves as the main feature-extracting module with three key components: a siamese semantic tokenizer, which groups pixels into concepts to generate a compact set of semantic tokens for each temporal input; a transformer encoder, which models context of semantic concepts in token-based space-time; and a siamese transformer decoder, which projects the corresponding semantic tokens back to pixel-space to obtain the refined feature map for each temporal image [26].

Let I_1, $I_2 \in R^{H\times W\times B}$ be the bitemporal images, where H, W, B is height, width, and number of bands of the image. BIT-based model first extracts high-level features X_1 and X_2 by a small CNN. Two token sets are computed by a semantic tokenizer based on the extracted features. Then, BIT models the global relationships within these token sets using a transformer encoder, resulting in context-rich semantic tokens T_1 and T_2. As the core of the transformer encoder, self-attention is calculated as follows:

$$\text{Attention}(Q,K,V) = \text{soft}\max\left(\frac{QK'}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are query, key, and value vectors from semantic tokens, respectively, and d_k represents the dimension of the key vector. It calculates the correlation weight matrix coefficients of Q and K and normalizes the weight matrix through softmax operation. The weight coefficients are superimposed on V to achieve modeling of global contextual information [26]. These context-enriched tokens contain high-level semantic details that effectively highlight the changes. To bridge the gap between these representations and pixel-level features, BIT employs a modified siamese transformer decoder to refine the image features for each image.

The final deep features obtained are f_1 and f_2, respectively. In summary, BIT interprets a feature map as a sequence of patches via its semantic tokenizer, facilitating the learning and correlation of global context related to high-level semantic concepts. The transformer's self-attention mechanism plays a crucial role in capturing long-range dependencies among pixels, enabling the modeling of comprehensive contextual information within images. As a result, BIT excels in comprehending the spatial relationships and overall characteristics of complex change targets in VHR images, substantially enhancing its capacity to represent change-related information.

C. Iteratively Reweighted Multivariate Alteration Detection

IR-MAD is an optimized iterative version of MAD, which essentially uses multivariate random variables to represent multispectral bitemporal images, and detects changes through multivariate statistical analysis [9]. The core step of IR-MAD is canonical correlation analysis, which involves deriving linear combinations from two sets of original variables and using correlation coefficients to analyze the correlation between two sets of variables, thereby reflecting the overall correlation between the original bitemporal images. The result obtained by calculating canonical variables and performing difference operations reflects the maximum change information of all spectral bands.

According to (2), a linear transformation is firstly performed on the input f_1 and f_2 using projection vectors a and b to obtain the U and V. Using U–V to represent the change information between images, the algorithm aims to find suitable a and b for maximizing the variance (3) of difference between U and V, that is, minimize the correlation between U and V, in order to concentrate as much change information as possible

$$\begin{cases} U = a^T f_1 \\ V = b^T f_2 \end{cases} \tag{2}$$

$$D(U-V) = 2(1-\text{Corr}(U,V)). \tag{3}$$

The variance-covariance matrix of f_1 and f_2 is (4). Using the Lagrangian multiplier method and denoting Corr (U, V) as r, (5) can be derived. The problem eventually turns into seeking eigenvalues r^2 and sorting them

$$\sum = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix} \tag{4}$$

$$\sum_{11}^{-1}\sum_{12}\sum_{22}^{-1}\sum_{21} a = r^2 a. \tag{5}$$

After finding the eigenvectors a and b, the corresponding canonical correlation variables can be calculated. The MAD variates (6) are linear combinations of input variables f_1 and f_2. The square sum of the MAD variates divided by the standard deviation approximately satisfies a chi-square distribution with p (number of bands) degrees of freedom. In addition, if there is no change at pixel j, then the ith MAD value, MADij, has a mean 0. On this basis, calculate the chi-square distance chi_j according to (7) and update the weight w_j of pixel according to (8). After the convergence of the canonical correlation coefficient, a thresholding algorithm is used to generate a binary CD map according to a predefined threshold value t.

Following the above principle, IR-MAD iteratively highlights change targets. It can effectively process multidimensional high-level features from the feature extraction module without excessive manual intervention

$$\text{MAD} = a^T f_1 - b^T f_2 \tag{6}$$

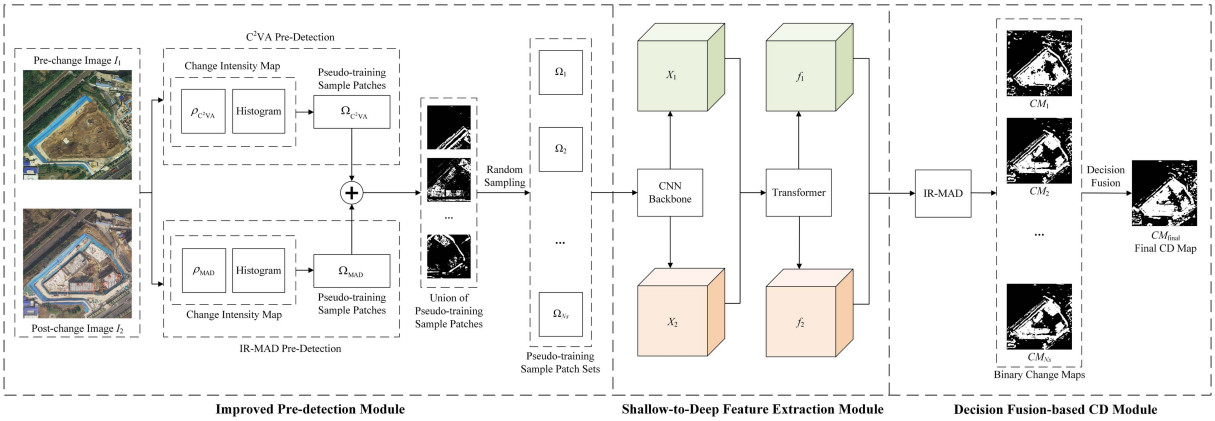

Fig. 1. Architecture of the proposed trans-MAD approach.

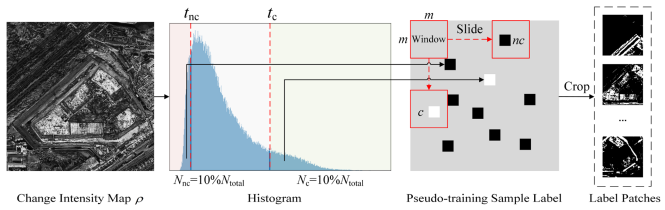

Fig. 2. Scheme of generating pseudotraining sample patches based on histogram of an intensity map.

$$\mathrm{chi}_j = \sum_{i=1}^{p} \left(\frac{\mathrm{MAD}_{ij}}{\sigma_{\mathrm{MAD}_i}} \right)^2 \in \chi^2(p) \tag{7}$$

$$w_j = P\{\mathrm{chi}_j > t\} = P\left\{\chi^2(p) > t\right\}. \tag{8}$$

III. Methodology

The proposed trans-MAD approach mainly consists of three modules: the improved predetection (IPD) module; the SD feature extraction module; and the DF-based CD module. The overall structure of the proposed approach is shown in Fig. 1. Details of each module are provided as follows.

A. Improved Predetection Module

This step aims to generate reliable pseudotraining samples for CD from the original image pair. To this end, C^{2}VA and IR-MAD methods are jointly considered to be used for pre-detection. Let I_1, $I_2 \in R^{H\times W\times B}$ be the bitemporal images with B bands. Two change intensity maps $\rho_{\mathrm{C}^2\mathrm{VA}}$ and ρ_{MAD} are computed independently according to the following equations:

$$\rho_{\mathrm{C}^2\mathrm{VA}} = \sqrt{\sum_{k=1}^{B} \left(I_1^k - I_2^k\right)^2} \tag{9}$$

$$\rho_{\mathrm{MAD}} = \sqrt{\sum_{k=1}^{B} \left(\frac{\mathrm{MAD}_k}{\sigma_{\mathrm{MAD}_k}}\right)^2}. \tag{10}$$

Fig. 2 shows the scheme of generating pseudotraining sample patches. Let N_{total} be the total number of image pixels. The number of no-change sample pixels (ω_{nc}) and change sample pixels (ω_c) are N_{nc} and N_c, respectively. Based on the intensity histogram, 10% N_{total} pixels with the lowest change intensity are selected as no-change pseudotraining samples, and 10% N_{total} pixels with the highest intensity are considered as change pseudotraining samples, while the remaining pixels are the background. The corresponding threshold values are t_{nc} and t_c for the no-change and change classes, respectively.

The sample pixel sets that generated by the C^{2}VA and IR-MAD pre-detection step are defined as $S_{\mathrm{C}^2\mathrm{VA}}$ and S_{MAD}, respectively. Then, a conflict elimination operation is performed. For a given pixel q in the image, its assigned categories in $S_{\mathrm{C}^2\mathrm{VA}}$ and S_{MAD} are $S_{\mathrm{C}^2\mathrm{VA}}$ and C_{MAD}, respectively. If q is selected as a sample pixel in both $S_{\mathrm{C}^2\mathrm{VA}}$ and S_{MAD}, but has category conflicts ($S_{\mathrm{C}^2\mathrm{VA}} \neq C_{\mathrm{MAD}}$), then it will be removed from $S_{\mathrm{C}^2\mathrm{VA}}$ and S_{MAD} simultaneously. This operation guarantees the unique and valid category of final obtained sample pixels.

After the above process, the sample pixel sets are reshaped into the shape of original image. Starting from the top left corner of the image, an $m\times m$ pixels window is used to slide and crop the image at a stride of s pixels. Finally, a series of sample patches containing pre-change, post-change images and binary labels were cropped out from two sets of pre-detection results, which are denoted as $\Omega_{\mathrm{C}^2\mathrm{VA}}$ and Ω_{MAD}, respectively.

In order to jointly utilize the information of two pre-detection algorithms to enhance the discriminative ability of the samples, we randomly select labels from two sample patch sets. In the nth sampling, corresponding to the previous cropping position (x, y), the final label L_{xy} will be selected from $\Omega_{\mathrm{C}^2\mathrm{VA}}$ or Ω_{MAD} with equal probability and added into final pseudotraining sample patch set Ω_n. This sampling step is independently repeated N_{s} (N_{s} usually is the odd value) times to obtain N_{s} pseudotraining sample patch sets $\{\Omega_1, \Omega_2, \ldots, \Omega_{N\mathrm{s}}\}$ for training deep learning models. The whole process of this step is summarized in Algorithm 1.

B. SD Feature Extraction Module

The purpose of this module is to extract advanced features based on the pseudotraining samples generated in the previous

Algorithm 1: Improved Pre-detection Module.

Input: Pre-change image I_1, Post-change image I_2
Output: Pseudo-training sample patch sets $\{\Omega_1, \Omega_2, \cdots, \Omega_{Ns}\}$
Obtain change intensity maps ρ_{C^2VA} and ρ_{MAD} according to I_1 and I_2;
Get the number of image pixels $N_{total} = H * W$;
Flatten the intensity map and sort it in ascending order;
Calculate the number of sample pixels $N_{nc} = N_c = 10\% * N_{total}$;
Select N_{nc} pixels with the lowest intensity and N_c pixels with the highest intensity as samples;
Obtain pseudo-training sample pixel sets S_{C^2VA} and S_{MAD};
Let q be a given pixel in the image;
if $q \in S_{C^2VA}$ *and* $q \in S_{MAD}$ **then**
  if $C_{C^2VA} \neq C_{MAD}$ **then**
    Remove q from S_{C^2VA} and S_{MAD} simultaneously;
Generate pseudo-training sample patch sets Ω_{C^2VA} and Ω_{MAD} by $m * m$ window clipping;
for $n = 1; n \leq N_s$ **do**
  for $x = 1; x \leq H - m$ **do**
    for $y = 1; y \leq W - m$ **do**
      Select final label L_{xy} from Ω_{C^2VA} or Ω_{MAD} with equal probability;
      Add L_{xy} into Ω_n;
final;
return $\{\Omega_1, \Omega_2, \cdots, \Omega_{Ns}\}$;

stage. This is realized by utilizing both CNN and BIT, which are complementary in feature extraction, thereby enriching the change representation information. In particular, CNN can capture the hidden change information within a local range of image through convolution operations, which is very helpful for identifying tiny differences in specific regions. Unlike this, BIT focuses more on global context and uses Transformer's self-attention mechanism to establish dependencies between distant pixels, thus enabling a better understanding of the overall structure and interrelationships of changed targets in the whole scene.

As shown in Fig. 2, a CNN backbone is firstly employed to process pseudotraining sample images from the predetection module and obtain the image features X_1 and X_2. Then, they are fed into BIT to generate refined global features f_1 and f_2. These deep features provide higher-level representation of original images and filter out some irrelevant information, enhancing the ability to capture complex changes that may not be easily distinguished in the original images. By taking advantages of the SD feature extraction process, both the local and global information of the bitemporal images is enhanced, which largely increases the change representability in the IR-MAD in the next step. By inputting the extracted SD features into the IR-MAD algorithm, it can effectively reduce the omission errors caused by directly extracting change information from the original bands.

Note that this feature extraction is also unsupervised and automatic relying on the generated pseudotraining samples, and can well model the local and global image information thus is more robust for CD, especially in dealing with the complex high-resolution images with limited spectral bands.

C. Decision Fusion-Based CD Module

This step is designed to generate the final CD results. Specifically, in the previous steps, a feature extraction framework combining CNN and transformer extracts SD features from the original pseudotraining sample images. These enhanced features are input into the CD process and utilized by IR-MAD algorithm to capture better change information. Unlike directly extracting change information from original spectral bands, this CD method reveals previously hidden information in the data, greatly improving the ability to detect and understand changes. The OTSU algorithm is used to generate the binary CD map. In addition, a majority voting fusion strategy is applied to the CD results obtained on different batches to improve the stability and reliability of the final output. Corresponding to the N_s sample sets $\{\Omega_1, \Omega_2, \ldots, \Omega_{Ns}\}$ obtained in the pre-detection process, a total of N_s CD maps $\{CM_1, CM_2, \ldots, CM_{Ns}\}$ were obtained. For a given pixel on the image, its final category C can be assigned according to the following rule:

$$C \in \begin{cases} \omega_c, \text{count} > \frac{N_s}{2} \\ \omega_{nc}, \text{otherwise} \end{cases} \tag{11}$$

where count represents the number of CD results that consider the pixel to be a change. Accordingly, the final binary CD map CM_{final} can be obtained.

By fusing several CD results through decision voting, commission errors can be effectively reduced to a certain extent, and the quality of the final output is better than that of a single detection CD map, with higher reliability.

(a) (b) (c)

Fig. 3. Bitemporal images of the GZ dataset obtained in (a) 2015, (b) 2017, and (c) change reference map.

IV. Dataset and Experimental Design

A. Dataset Description

Two satellite remote sensing VHR images CD datasets are considered to evaluate the proposed approach in the experiment, which are introduced as follows.

1) Guangzhou (GZ) Dataset [34]: This is a large-scale VHR multispectral satellite image data set that acquired by Google Earth service, covering the suburb of Guangzhou, China. RGB color images were considered with a spatial resolution of 0.55 m. One pair of images (1836 × 1836 pixels) was selected in our experiment, where the main changes are related to the buildings, farmland, bare land, etc. in the image scenario (denoted as GZ dataset). Note that the initial annotation only focuses on the specific change of buildings, thus we carefully modified the original change map and carried out further labeling to add more comprehensive change classes including farmland, bare land, waters, and roads. The considered VHR images and the corresponding change reference map are illustrated in Fig. 3.

2) Nanjing (NJ) Dataset: This dataset is made up of two pan-sharpened multispectral images acquired by BJ-3 satellite in 2022 and 2023, covering the urban area of Nanjing city, which

(a) (b) (c)

Fig. 4. Bitemporal images of the NJ dataset obtained in (a) 2022, (b) 2023, and (c) change reference map.

is denoted as the NJ data set. The pair of bitemporal images consist of three spectral bands (red, green and blue) with a spatial resolution of 0.5 m and 512 × 512 pixels. A change reference map [see Fig. 4(c)] was made by careful image interpretation, and changes in this scene were mainly buildings, vegetation, road, and bare land.

B. Experimental Design

In order to evaluate the effectiveness of the proposed trans-MAD approach, several SOTA unsupervised CD methods are considered for a comparison purpose, including four traditional algorithms, i.e., C^{2}VA [6], SFA [11], MAD [8], IR-MAD [9], and three deep learning-based CD methods, i.e., DCVA [28], DSFA [30], and KPCA-MNet [31].

The proposed algorithm was implemented using the PyTorch framework on an NVIDIA RTX2050 GPU with 4GB memory. The basic learning rate was set to 0.01 and decreased linearly. The SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4} was applied to update the learning rate based on the training data and improve convergence speed. The batch size was set as 4 according to the GPU memory. The size of the cropping window m was 256 when generating pseudotraining sample patches. The sliding stride s was set to 64 and 16 for the GZ and NJ datasets, respectively.

Parameter settings for the reference methods were as follows. Layers in DCVA was set as $\{2, 5, 8\}$. The regularization parameter in DSFA was 10^{-4}, and the network had 2 hidden layers with 128 nodes per layer. The network depth of KPCA-MNet and the number of KPCA convolution kernels were 4 and 8, respectively. The radial basis function kernel was chosen as the kernel function, with kernel parameter equal to 5×10^{-4}. The convolution kernel's size was set to 3 according to the experimental performance.

C. Evaluation Metrics

In binary change maps, the changed (positive) areas are represented by white pixels, while the unchanged (negative) regions are represented by black pixels. True positive (TP) indicates the number of changed pixels which are correctly detected. False positive (FP) is the number of unchanged pixels that are falsely detected as changed ones. True negative (TN) means the number of pixels predicted correctly as unchanged, and false negative (FN) denotes the number of pixels predicted incorrectly as unchanged.

TABLE I
ACCURACY INDICES ON THE GZ DATASET

Methods	OA	KC	Precision	Recall	F1
C^{2}VA	0.7046	0.3629	0.8583	0.3923	0.5385
SFA	0.6324	0.2337	0.6078	0.4602	0.5238
MAD	0.6539	0.2820	0.6322	0.5071	0.5628
IR-MAD	0.6669	0.3008	0.6741	0.4683	0.5526
DCVA	0.6252	0.1928	0.6562	0.3084	0.4196
DSFA	0.5897	0.1137	0.5741	0.2552	0.3534
KPCA-MNet	0.4876	-0.0079	0.4358	0.5644	0.4918
Trans-MAD	**0.8116**	**0.6088**	**0.8592**	**0.6830**	**0.7610**

Quantitative evaluation was carried out based on the obtained binary CD maps and the corresponding pixel-level change reference map. To this end, overall accuracy (OA), Kappa coefficient (KC), precision, recall, and F1-score (F1) were calculated by the following formulas [32]:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{12}$$

$$\mathrm{KC} = \frac{\mathrm{OA} - \mathrm{PE}}{1 - \mathrm{PE}} \tag{13}$$

$$\mathrm{PE} = \frac{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN}) + (\mathrm{FN}+\mathrm{TN})(\mathrm{FP}+\mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})^2} \tag{14}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{15}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{16}$$

$$F1 = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{17}$$

In particular, OA represents the proportion of the number of correctly classified samples to the total number of samples. KC describes the similarity between the CD result and the ground truth. Precision can represent the proportion of TP pixels in the number of pixels detected as positive (changed). Recall represents the proportion of TP pixels in the number of pixels that are actually positive (changed) on the ground. F1 is the harmonic average calculated by the comprehensive Precision rate and recall rate, which represents the level of precision and recall at the same time.

V. RESULTS AND DISCUSSION

A. Performance Comparison

The proposed approach and other seven reference methods were tested on two considered VHR-CD data sets. Obtained results are shown in Figs. 5 and 6, and Tables I and II. The results in the second row of Figs. 5 and 6 are locally enlarged images, with enlarged areas marked in red boxes in Figs. 3 and 4, respectively.

In particular, for the GZ data, as shown in Fig. 5, one can see that compared with the reference methods, the proposed trans-MAD resulted in the best binary CD map with lower omission

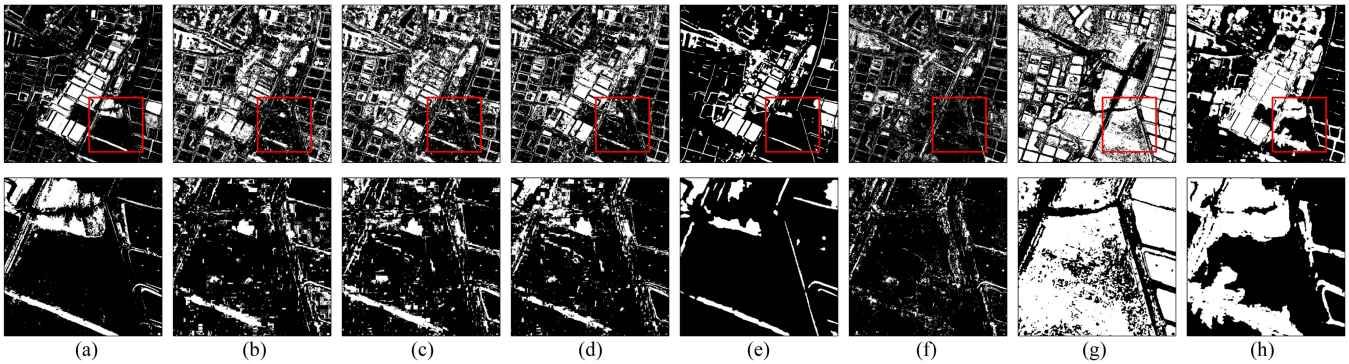

Fig. 5. Binary change maps obtained by different methods on the GZ dataset. (a) Compressed change vector analysis. (b) Slow feature analysis. (c) Multivariate alteration detection. (d) Iteratively reweighted multivariate alteration detection. (e) Deep change vector analysis. (f) Deep SFA. (g) Kernel principal component analysis convolution mapping network. (h) Trans-MAD. First row: Whole CD maps; second row: Enlarged subsets of the whole CD maps as highlighted in red box.

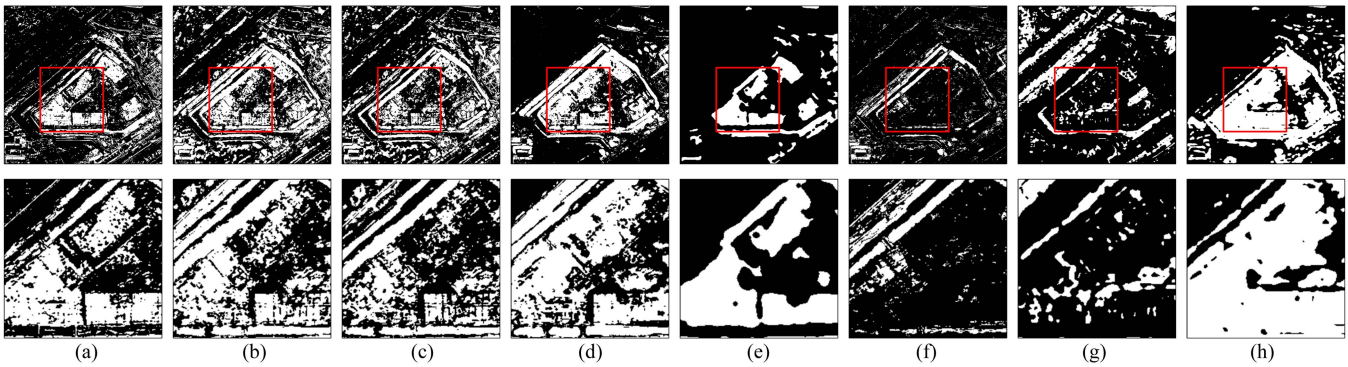

Fig. 6. Binary change maps obtained by different methods on the NJ dataset. (a) Compressed change vector analysis. (b) Slow feature analysis. (c) Multivariate alteration detection. (d) Iteratively reweighted multivariate alteration detection. (e) Deep change vector analysis. (f) Deep SFA. (g) Kernel principal component analysis convolution mapping network. (h) Trans-MAD. First row: Whole CD maps; second row: Enlarged subsets of the whole CD maps as highlighted in red box.

TABLE II
ACCURACY INDICES ON THE NJ DATASET

Methods	OA	KC	Precision	Recall	F1
C^2VA	0.7549	0.4253	0.7290	0.4979	0.5917
SFA	0.7082	0.3651	0.5903	0.5945	0.5924
MAD	0.7152	0.3623	0.6144	0.5409	0.5753
IR-MAD	0.7680	0.4595	0.7459	0.5299	0.6196
DCVA	0.7571	0.3945	**0.8672**	0.3768	0.5253
DSFA	0.6685	0.1432	0.6193	0.1831	0.2827
KPCA-MNet	0.5344	-0.1255	0.2261	0.1260	0.1619
Trans-MAD	**0.7753**	**0.5010**	0.7022	**0.6427**	**0.6711**

errors, fewer noises, and a more regular and well-recognized changed region. This is benefit from the following aspects: it effectively generates samples representing various types of changes by the designed IPD strategy. Then, an advanced SD feature extraction module is utilized to better capture features in VHR images, and local and global information about changes is modeled to effectively reduce omission errors. The DF output step also plays an important role in reducing the uncertainty of pseudotraining sample generation, reducing false alarms and making the final CD result more accurate. From Fig. 5(h), it is worth noting that trans-MAD effectively identified changes with varying significance linked to different land-cover transitions (e.g., buildings, vegetation, and bare land changes). However, the performance of KPCA-MNet was relatively poor due to a large number of commissions. This may be caused by the limited training samples and insufficient representativeness of features extracted by KPCA convolution, and by its incompatibility to the specific scenario, finally leading to false alarms occurred in the detected soil and water changes. The CD map that produced by DCVA exhibited fewer salt-and-pepper noises due to the morphological processing. However, it produced many omissions associated to the vegetation changes. This could be attributed to the limitation of the feature extraction ability of the pretrained model used in the algorithm. As for the DSFA method, its detection result was fragmented with many omission errors, which may be affected by sample selection limitations and the poor feature extraction in fully connected networks. However, as we can see from the results, for the compared deep learning-based CD reference methods, they only focus on the most significant land-cover changes in the scene (i.e., mainly are building changes). For the reference four traditional unsupervised CD methods, many commission errors were presented in the obtained CD maps. The main reason is that they only

utilize the shallow spectral features (i.e., original bands) of VHR images without considering more robust deep features.

Table I gives the quantitative evaluation indices obtained by different CD methods on GZ data. The best values of each evaluation criteria are highlighted in bold. As one can see, the proposed trans-MAD achieved significantly higher accuracies than other methods on all criteria. In particular, the OA value (i.e., 0.8116) was 18.64% higher than the best in deep learning methods and 10.7% higher than the best in traditional methods. KC value (i.e., 0.6088) was 41.6% and 24.59% higher than the best performance of the deep learning method and traditional method, respectively. In addition, the precision, recall, and F1 of the proposed approach were 0.8592, 0.6830, and 0.7610, respectively, also ranking first among the comparison methods. A higher Precision rate demonstrates that the proposed trans-MAD approach presents fewer commission errors in CD. In the meantime, excellent performance on recall values indicates that fewer change areas are falsely detected as unchanged by trans-MAD. In summary, among all traditional CD methods, C^2VA performs best on OA, KC, and Precision indicators, while MAD-based methods perform better on Recall and F1. In deep learning methods, the proposed trans-MAD achieves the best performance.

For the NJ dataset, the binary CD maps obtained by the proposed trans-MAD approach and the reference methods are given in Table II. Compared with other reference methods, the proposed trans-MAD obtained the most accurate CD map, and comprehensive qualitative evaluation showed its superior performance. Specifically, the result of the trans-MAD method [see Fig. 6(h)] showed significant noise suppression compared to the CD results of the traditional SFA method [see Fig. 6(b)] and MAD method [see Fig. 6(c)], effectively avoiding the creation of many unrelated error detection regions. Compared with the CD method based on deep learning [see Fig. 6(e)–(g)], the trans-MAD approach had the smallest number of omission errors and significant advantage in generating more regular and meaningful change maps. Experimental results confirm that trans-MAD has the ability to capture both significant and insignificant changes in VHR images with remarkable clarity.

According to the statistical results given in Table II, the proposed trans-MAD outperformed the reference algorithms on OA values (i.e., 0.7753), which indicated that more change and unchanged areas were accurately detected. The KC value (i.e., 0.5010) was 4.15% higher than the second highest value from IR-MAD, indicating a higher consistency between the CD output of trans-MAD and the reference change map. The recall value (i.e., 0.6427) and F1 value (i.e., 0.6711) of trans-MAD exceeded the second highest scoring method by 4.82% and 5.15%, respectively. The advantage of the trans-MAD approach in Recall indicates that it has fewer omission errors than other reference algorithms. The leading performance of the trans-MAD method in F1-score demonstrates that the algorithm strikes a good balance between omission and commission errors, thus can achieve high Precision and Recall values simultaneously. The quantitative results confirm that by considering the trade-off between accuracies and CD errors (i.e., omissions and commissions), the proposed trans-MAD achieves overall excellent CD performance with the best OA, KC, Recall, and F1 values.

TABLE III
ABLATION ANALYSIS ON DIFFERENT MODULES ON THE GZ DATASET

Methods	OA	KC	Precision	Recall	F1
Trans-MAD	**0.8116**	**0.6088**	0.8592	**0.6830**	**0.7610**
w/o IPD (C^2VA)	0.6741	0.2847	**0.9199**	0.2826	0.4324
w/o IPD (IR-MAD)	0.8034	0.5913	0.8516	0.6689	0.7493
w/o DF	0.7887	0.5634	0.8096	0.6786	0.7383

TABLE IV
ABLATION ANALYSIS ON DIFFERENT MODULES ON THE NJ DATASET

Methods	OA	KC	Precision	Recall	F1
Trans-MAD	**0.7753**	**0.5010**	0.7022	**0.6427**	**0.6711**
w/o IPD (C^2VA)	0.7286	0.3919	0.6365	0.5575	0.5944
w/o IPD (IR-MAD)	0.7601	0.4266	**0.7714**	0.4653	0.5805
w/o DF	0.6721	0.2551	0.5493	0.4491	0.4942

B. Ablation Analysis

In this section, ablation experiments were carried out on GZ and NJ dataset to evaluate the individual performance and contribution of newly-added modules in the proposed trans-MAD model. We evaluated the significance of IPD module and DF module in trans-MAD model. Unless otherwise indicated, all experimental settings remain consistent and comparable. The findings are given in Tables III and IV .

1) *IPD Module*: This module is a core component in trans-MAD responsible for generating pseudotraining samples. It utilizes C^2VA and IR-MAD algorithm simultaneously to recognize the basic change areas with high confidence from bitemporal images. To determine the effectiveness of IPD module in generating pseudotraining samples, we replaced it with a single pre-detection algorithm (i.e., C^2VA or IR-MAD). As one can see from the obtained experiment results, compared with the trans-MAD, the predetection module with only C^2VA led to a significant decrease of recall values from 0.6830 to 0.2826 on GZ dataset, and from 0.6427 to 0.5575 on NJ dataset. F1-scores also decreased over 32% and 7%, respectively, on two datasets. This highlights the importance of the IPD module. Additional experiments using IR-MAD-only predetection showed the same trend as well. In summary, the superiority of change information capturing and pseudotraining samples generation ability using an IPD module is validated.
2) *Decision Fusion-based CD Module*: We also performed an ablation analysis to evaluate the DF Module in the trans-MAD. As given in Tables III and IV, methods without the DF component (w/o DF) exhibited poor performance. On the GZ dataset, OA values decreased from 0.8116 to 0.7887, Precision values dropped from 0.8592 to 0.8096, and F1-scores declined from 0.7610 to 0.7383. On the NJ dataset, the OA, precision and F1 values reduced from 0.7753, 0.7022, and 0.6711 to 0.6721, 0.5493, and 0.4942, respectively. This suggests the DF module can effectively reduce CD errors to improve the final CD accuracy, especially the commission errors.

TABLE V
NOTATIONS USED IN THIS ARTICLE

Symbol	Description
I	Input image
H	Height of image
W	Width of image
B	Number of bands of image
X	Feature extracted by CNN
T	Semantic Token
Q	Query vector
K	Key vector
V	Value vector
d_k	Dimension of key
f	Feature extracted by Transformer
a and b	Projection vectors used in IR-MAD algorithm
U and V	Linear transformations of original image
$D\,(U, V)$	Variance of (U, V)
r	Correlation coefficient
Σ	Variance-covariance matrix
p	Degree of freedom of Chi-square distribution
MAD_{ij}	The ith MAD value of pixel j
σ	Standard deviation
chi	Chi-square distance
w_j	Weight of pixel j
P	Probability
ρ	Intensity of change
N_{total}	Number of pixels of an input image
N_{nc}	Number of no-change sample pixels
N_{c}	Number of change sample pixels
t_{nc}	Threshold for determining no-change pixels
t_{c}	Threshold for determining change pixels
ω_{nc}	No-change sample pixels
ω_{c}	Change sample pixels
S	Pseudo-training sample pixels set
C	Category of pixel
m	Size of the cropping window
s	Stride of the cropping window
Ω	Pseudo-training sample patches set
L_{xy}	Label patch at position (x, y)
N_{s}	Number of sample sets
CM	Change detection map
$count$	Number of CD maps that consider the pixel to be a change

VI. CONCLUSION

In this article, a new unsupervised deep learning CD method named trans-MAD was developed. In order to achieve unsupervised network training, it performs an IPD by taking advantages of C^{2}VA and IR-MAD, enhancing the quality and diversity of generated pseudotraining samples. A deep learning-based SD feature extraction scheme is utilized relying on CNN and Transformer and integrated with the IR-MAD algorithm. It achieves more effective feature expression, enhances change representation, as well as reduces CD omissions. In addition, a DF-based CD step is implemented to reduce commission errors, thus improving the overall accuracy of CD. The proposed approach exhibits excellent performance in complex urban scenarios, which is superior to other advanced unsupervised methods. Note that it can better identify diverse types of changes with different significance levels in the scene, rather than only focusing on the most significant changes (e.g., buildings).

Experiment results obtained on two VHR remote sensing CD datasets confirm the effectiveness of the proposed approach in identifying accurately complex changes over different scenes. Comprehensive qualitative and quantitative evaluations demonstrated that the proposed trans-MAD method outperforms other compared methods, showing the potential in land-cover CD. Future research will be devoted to improving the adaptability and automation of the predetection algorithm. In addition, the possibility of extending the ability of trans-MAD on multiclass CD is worth further study.

REFERENCES

[1] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019, doi: 10.1109/MGRS.2019.2898520.

[2] S. Liu, Q. Du, X. Tong, A. Samat, and L. Bruzzone, "Unsupervised change detection in multispectral remote sensing images via spectral-spatial band expansion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3578–3587, Sep. 2019, doi: 10.1109/JSTARS.2019.2929514.

[3] S. Liu, F. Bovolo, L. Bruzzone, X. Tong, and Q. Du, "Editorial foreword to the special issue on recent advances in multitemporal remote-sensing data processing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 776–778, 2022, doi: 10.1109/JSTARS.2022.3140594.

[4] D. Lu, P. Mausel, E. Brondízio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, Jun. 2004.

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

[6] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2196–2212, Jun. 2012, doi: 10.1109/TGRS.2011.2171493.

[7] S. Liu, L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du, "Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4363–4378, Aug. 2015, doi: 10.1109/TGRS.2015.2396686.

[8] 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, Apr. 1998.

[9] 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, doi: 10.1109/TIP.2006.888195.

[10] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009, doi: 10.1109/LGRS.2009.2025059.

[11] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014, doi: 10.1109/TGRS.2013.2266673.

[12] S. Liu, Q. Du, X. Tong, A. Samat, L. Bruzzone, and F. Bovolo, "Multiscale morphological compressed change vector analysis for unsupervised multiple change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4124–4137, Sep. 2017, doi: 10.1109/JSTARS.2017.2712119.

[13] Y. Feng, S. Liu, and L. Tang, "Automatic extraction and change monitoring of fire disaster event based on high-resolution nighttime light remote sensing images," in *Proc. Image Signal Process. Remote Sens. 26th*, 2020, pp. 57–65.

[14] 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, doi: 10.1109/TGRS.2015.2505183.

[15] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019, doi: 10.1109/TIP.2018.2878958.

[16] Y. Zheng, S. Liu, Q. Du, H. Zhao, X. Tong, and M. Dalponte, "A novel multitemporal deep fusion network (MDFN) for short-term multitemporal HR images classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10691–10704, 2021, doi: 10.1109/JSTARS.2021.3119942.

[17] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021, doi: 10.1109/TGRS.2020.3016820.

[18] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010, doi: 10.1109/TGRS.2021.3124913.

[19] S. Liu, H. Zhao, Q. Du, L. Bruzzone, A. Samat, and X. Tong, "Novel cross-resolution feature-level fusion for joint classification of multispectral and panchromatic remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619314, doi: 10.1109/TGRS.2021.3127710.

[20] S. Liu et al., "A shallow-to-deep feature fusion network for VHR remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410213, doi: 10.1109/TGRS.2022.3179288.

[21] H. Zhao et al., "GCFnet: Global collaborative fusion network for multispectral and panchromatic image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5632814, doi: 10.1109/TGRS.2022.3215020.

[22] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412, doi: 10.1109/TGRS.2023.3279834.

[23] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856, doi: 10.1016/j.rse.2023.113856.

[24] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067, doi: 10.1109/ICIP.2018.8451652.

[25] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, Jun. 2019, Art. no. 1382.

[26] 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, doi: 10.1109/TGRS.2021.3095166.

[27] M. Gong, X. Niu, P. Zhang, and Z. Li, "Generative adversarial networks for change detection in multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2310–2314, Dec. 2017, doi: 10.1109/LGRS.2017.2762694.

[28] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019, doi: 10.1109/TGRS.2018.2886643.

[29] H. Chen, C. Wu, B. Du, and L. Zhang, "Deep siamese multi-scale convolutional network for change detection in multi-temporal VHR images," in *Proc. 10th Int. Workshop Anal. Multitemporal Remote Sens. Images*, 2019, pp. 1–4, doi: 10.1109/Multi-Temp.2019.8866947.

[30] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019, doi: 10.1109/TGRS.2019.2930682.

[31] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12084–12098, Nov. 2022, doi: 10.1109/TCYB.2021.3086884.

[32] Q. Ke and P. Zhang, "Hybrid-TransCD: A hybrid transformer remote sensing image change detection network via token aggregation," *Int. Soc. Photogramm. Remote Sens. Int. J. Geo-Inf.*, vol. 11, no. 4, Apr. 2022, Art. no. 263, doi: 10.3390/ijgi11040263.

[33] 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, doi: 10.1109/TGRS.2022.3169479.

[34] 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, doi: 10.1109/TGRS.2020.3011913.

Yizhang Lin received the B.E. degree in surveying and mapping engineering in 2022 from Tongji University, Shanghai, China, where he is currently working toward the M.E. degree in surveying and mapping science and technology.

His current research interests include multispectral image change detection, and in-situ spectral analysis of Martian mineral composition.

Sicong Liu (Senior Member, IEEE) received the B.Sc. degree in geographical information system and the M.E. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2009 and 2011, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, 2015.

He is currently an Associate Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. His research interests include multitemporal data analysis, multispectral/hyperspectral remote sensing in earth observation and planetary exploration.

Dr. Liu was the recipient (ranked as third place) of Paper Contest of the 2014 IEEE GRSS Data Fusion Contest. He is the Technical Co-Chair of the Tenth International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp 2019). He is the Program Committee Member for SPIE Remote Sensing Symposium: Image and Signal Processing for Remote Sensing (since 2020), and was the Session Chair for many international conferences such as International Geoscience and Remote Sensing Symposium (since 2017). He is an Associate Editor and a Guest Editors for several international journals.

Yongjie Zheng (Student Member, IEEE) received the B.S. degree in remote sensing science and technology from Henan Polytechnic University, Jiaozuo, China, in 2018, and the M.S. degree in photogrammetry and remote sensing from Tongji University, Shanghai, China, in 2021. She is currently working toward the Ph.D. degree in remote sensing image analysis with the University of Trento, Trento, Italy.

Her research interests include deep learning, feature extraction, feature fusion, and remote sensing image classification and change detection.

Xiaohua Tong (Senior Member, IEEE) received the Ph.D. degree in traffic engineering from Tongji University, Shanghai, China, in 1999. He was a Postdoctoral Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, between 2001 and 2003.

He was a Research Fellow with the Hong Kong Polytechnic University in 2006, and a Visiting Scholar with the University of California, Santa Barbara, CA, USA, between 2008 and 2009. His current research interests include remote sensing, geographic information system, uncertainty and spatial data quality, image processing for high resolution, and hyperspectral images.

Dr. Tong is the Vice-Chair of the Commission on Spatial Data Quality of the International Cartographical Association, and the Co-Chair of the ISPRS working group (WG II/4) on Spatial Statistics and Uncertainty Modeling.

Huan Xie (Senior Member, IEEE) received the B.S. degree in surveying engineering, and the M.S. and Ph.D. degrees in cartography and geoinfomation from Tongji University, Shanghai, China, in 2003, 2006, and 2009, respectively.

From 2007 to 2008, she was a Visiting Scholar with the Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany, funded by the China Scholarship Council. Her research interests include satellite laser altimetry and hyperspectral remote sensing.

Hongming Zhu (Member , IEEE) received the B.S. degree in computer science, and the M.S. degree in computer architecture from Tongji University, Shanghai, China, in 2002, and 2005, respectively, and the Ph.D. degree in information and communication technology from the University of Bolton, Bolton, U.K., 2017.

He is currently an Associate Professor with the School of Software Engineering Tongji University, Shanghai, China. His research interests include multimodality data fusion, object detection and change detection, hyperspectral images processing.

Kecheng Du received the B.S. degree in surveying and mapping science from Wuhan University, Wuhan, China, in 2020. He is currently working toward the M.S. degree in surveying and mapping with Tongji University, Shanghai, China.

His current research interests include hyperspectral image change detection, lunar in situ spectral analysis and mineral content inversion.

Hui Zhao received the B.S. degree in geomatics engineering from Chuzhou University, Chuzhou, China, in 2012, and the M.S. degree in geomatics engineering from Jiangxi University of Science and Technology, Ganzhou, China, in 2017. She is currently working toward the Ph.D. degree in surveying and mapping with Tongji University, Shanghai, China.

Her current research interests include multi-source remote sensing data fusion in the field of earth observation and planetary exploration.

Jie Zhang received the B.S. degree in geography and environmental resources from Shandong Normal University, Jinan, China, in 2019, and the M.S. degree in cartography and geographical information system from Capital Normal University, Beijing, China, in 2022. She is currently working toward the Ph.D. degree in surveying and mapping with Tongji University, Shanghai, China.

Her current research interests include Mars topography mapping and planetary exploration.