A Spatial–Temporal Difference Aggregation Network for Gaofen-2 Multitemporal Image in Cropland Change Area

Chuang Liu, Liyang Bao, and Zhiqi Zhang

***Abstract*—Food security is an important guarantee of peace and development in the world. The accurate monitoring of cropland utilizing remote sensing data provides a strong technical support for the protection of cropland resources. Nonetheless, in contrast to the building change detection, the growth characteristics of crops in cropland areas exhibit significant variations in accordance with different seasonal climates and light intensities. Furthermore, the serious imbalance between the cropland change area and the nonchange area makes it difficult to focus on the real change area in the cropland under these interferences. To this end, we propose a spatial–temporal difference aggregation network (STDAN) for cropland change detection (CCD), which can focus on the real change area between different temporal images. Specifically, we use a cross-temporal difference feature enhancement module to enhance the difference features while establishing the correlation between different temporal features, which can suppress task-independent interference. Subsequently, the cross-level difference feature aggregation (CDFA) realizes the aggregation between different levels of difference features in an incremental manner to further refine the change area. Finally, the utilization of multireceptive fusion enables the integration of different scale characteristics obtained by CDFA, thereby yielding the accurate CCD outcomes. The experimental results indicate that the proposed STDAN achieves the highest *F*1, IOU, OA, and Kappa scores at 79.63%, 66.16%, 97.05%, and 78.04%, respectively, on the Gaofen-2 cropland data. In addition, we conduct generalization experiments on the remaining three mainstream datasets, demonstrating that our method is equally applicable to other change detection scenarios.**

***Index Terms*—Cropland change detection (CCD), Gaofen-2, multispectral image, multitemporal image, remote sensing.**

I. Introduction

THE concept of food security is fundamental to the peaceful development of human society, and cropland as the main producing area of human food holds unparalleled significance [1], [2]. Nowadays, with the growth of population and the expansion of cities, the quantity and quality of cropland have sharply declined, posing a significant threat to food security. Fortunately, the rapid development of modern satellites and sensors provides strong data support for real-world application. For example, the Gaofen-2 satellite from China can provide high-resolution images for the efficient monitoring of real-time changes in cropland areas [3], [4]. These data have the characteristics of high quality, high spatial resolution, and high temporal resolution. These high spatial resolution multitemporal remote sensing data can be utilized to detect changes in the cropland area in a timely and accurate manner. Hence, researchers endeavor to employ change detection technology to analyze different temporal remote sensing data and furnish technical assistance for the continuous monitoring of cropland areas [5]. The existing cropland change detection (CCD) methods are mainly divided into traditional methods, machine learning (ML) based methods, and deep learning (DL) based methods.

Algebraic-based and transform-based methods are the two types of traditional methods. Algebra-based methods primarily analyze the intensity and direction of each pixel in multitemporal images through mathematical operations. Common mathematical operations include image difference and image ratio [6], [7]. The difference operation is capable of extracting the change information between different temporal images. However, this simple operation is incapable of distinguishing the variations in brightness, rendering it susceptible to factors, such as illumination conditions. Another algebra-based method extracts difference features by calculating the pixel ratio between bitemporal images. Compared with the difference operation, this method is better able to distinguish variations in brightness. However, this method necessitates manually setting the threshold to determine whether the pixel undergoes changes, rendering it less robust to noise. The transformation-based methods aim to model different temporal images through transformation operations to extract the difference features. Commonly used transformation analysis tools include principal component analysis (PCA) [8], tasseled cap transformation (TCT) [9], and change vector analysis (CVA) [10]. PCA is a linear dimensionality reduction technique that searches for the main change direction of the data by calculating the covariance matrix of the data, which is the principal component, in order to extract the change characteristics. PCA is capable of identifying the most significant change patterns in different temporal images. The TCT algorithm is derived from morphology, which permits the extraction of edge

Received 5 September 2024; revised 9 December 2024; accepted 21 December 2024. Date of publication 24 December 2024; date of current version 10 January 2025. This work was supported by the National Key R&D Program of China under Grant 2022YFB3902800. *Chuang Liu and Liyang Bao contributed equally to this work. (Corresponding author: Zhiqi Zhang.)*

Chuang Liu and Liyang Bao are with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China (e-mail: liuchuang@hbut.edu.cn; 102211123@hbut.edu.cn).

Zhiqi Zhang is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zzq540@hbut.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3522066

texture characteristics through brightness differences between bitemporal images. CVA can determine whether there is a change by analyzing the size and direction of the pixel-level change vector, similar to the image ratio. The detection performance of this type of method is highly dependent on the precision of the manually set threshold, and it is susceptible to noise, such as illumination, resulting in a high rate of false detection and missed detection in the application of CCD.

The above-mentioned traditional method is difficult to apply to the actual cropland change scene, given the increasing quality of the remote sensing images captured by Gaofen-2 sensors. In order to improve the accuracy of CCD, some researchers consider it to be a binary classification problem and employ ML technology to address this problem. The commonly employed classifiers comprise of support vector machine (SVM) [11], random forest (RF) [12], and k-nearest neighbor algorithm (KNN) [13]. SVM possesses the capability to locate the optimal plane in high-dimensional space, thereby enabling it to accomplish the classification task of small samples in intricate scenes. However, the computational complexity of SVM makes it difficult to match the actual application requirements when the amount of data grows. RF trains multiple decision trees by randomly selecting features and samples to enable them to obtain more robust classification results. In contrast to RF, the implementation process of KNN is more straightforward, and it exhibits a certain degree of robustness toward outliers and noise. Nonetheless, as the identification of different temporal images necessitates a considerable amount of computation, these classification algorithms are less efficacious when applied to CCD tasks. In addition, similar to pixel-based methods, ML-based methods still require the manual setting of parameters and attributes, resulting in uneven performance across different datasets.

In recent years, with the improvement of the quality and quantity of remote sensing data and the ability of computer computing, the method based on DL has been widely used in the field of CCD due to its strong nonlinear fitting ability [14], [15], [16]. Inspired by the semantic segmentation task [17], some UNet-based single-branch CCD networks, such as FC-EF [18] and FresUNet [19], have been proposed. These methods are capable of locating the change area and obtaining the semantic information about the change area. With the in-depth study of change detection tasks, many researchers have used the idea of Siamese networks to design CCD models. Based on the framework of Siamese network, Daudt et al. [18] directly calculated the absolute value of the same scale features in the encoder and predicted the change area according to the series of the differences at all levels. Zhang et al. [20] proposed a deep supervised Siamese network, which can integrate heterogeneous features from both the channel and the spatial dimensions. Fang et al. [21] proposed a dense skip-connected Siamese network, which combines multiscale features and can reduce the local information loss caused by the excessive number of layers of the neural network. Jiang et al. [22] proposed a multiscale Siamese network to extract multiscale information from different temporal images. Since self-attention can establish global dependencies, some networks, such as BIT [23], WNet [28], and EATDer [24], use transformers to model different temporal images to capture global spatial–temporal information. Xie et al. [25] proposed WBANet, which combines the wavelet transform with self-attention to preserve high-frequency information during downsampling. In addition, cross attention is widely employed in the field of CCD, which can establish semantic relationships between different sequences. In order to enhance the extraction of difference features, Wu et al. [26] proposed cross Swin transformer, which is capable of interacting feature maps in two temporal branches with identical spatial resolution. To establish local and global dependencies simultaneously, several hybrid convolution and transformer methods have been proposed. Zhang et al. [27] developed parallel convolution and self-attention modules to capture global semantic information. Tang et al. [28] employed the convolutional neural network (CNN) and transformer to extract both local information and global correlations simultaneously, thereby capturing irregularly changing regions with greater precision. Meanwhile, some methods focus on tackling pseudochange and category imbalance. A graph-based knowledge supplementation network, proposed by Wang et al. [29], is capable of handling the errors in pseudolabeled samples, which in turn minimizes the detrimental effects of noisy samples. Hang et al. [30] propose AANet, which calculates the respective weight occupancy based on the scale difference between different change regions and performs difference fusion to alleviate the pseudochange problem in change detection. Liu et al. [31] introduce graph convolution to model spatial information in diverse temporal images and employ the attention mechanism to harmonize spatial and spectral information. This way, they can suppress the influence of unchanged regions. Feng et al. [32] utilized the complementary advantages of self-attention and cross attention to interact before differentiating features to learn the global distribution of each input feature. This approach mitigates category imbalance, thereby achieving the accurate localization of changing area.

In general, DL-based methods exhibit superior robustness and generalization ability. However, in the CCD scenario, there are still three limitations that affect the detection performance of DL-based methods.

1) As shown in Fig. 1(a), in addition to the light factor, the vegetation in cropland is greatly influenced by seasons. This is primarily due to the fact that vegetation may exhibit fundamentally different growth characteristics during different growth periods.
2) As shown in Fig. 1(b), illegal constructions in cropland tend not to expand greatly and their proportion is relatively small. This further exacerbates the imbalance between the categories of nonchanging cropland areas and changing areas.
3) In the CCD task, the two situations described above can exacerbate each other. The changing building area exhibits identical characteristics to the nonchanging cropland area, owing to the influence of pseudochanges, which can further exacerbate missed detection and misdetection.

In order to accurately detect the real change area in cropland, we propose a spatial–temporal difference aggregation network (STDAN) for CCD. Initially, STDAN enhances the difference

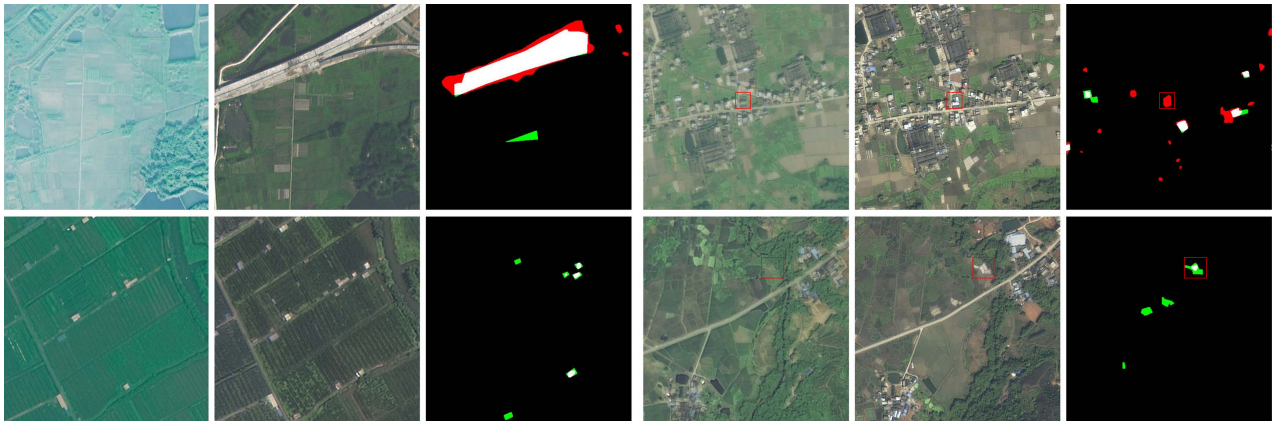

Fig. 1. Two limitations that affect the detection performance of DL-based methods in CCD task. (a) Some pseudochanges, such as seasonal changes and light intensity. (b) Imbalance between the nonchanging area and the changing area.

features while establishing the correlation between different temporal features using the cross-temporal difference feature enhancement module (TDE), which can suppress task-independent interference. Subsequently, the cross-level difference feature aggregation (CDFA) enables the aggregation between different levels of difference features in an incremental manner, thereby further refining the change area. Ultimately, the accurate results can be obtained by integrating the different characteristics at different scales in CDFA through multireceptive fusion (MRF). In contrast to the related methods, STDAN is capable of simultaneously taking into account the differences in crop characteristics resulting from pseudochanges, the similarity between changing and nonchanging cropland regions resulting from category imbalance in the CCD scene, and the interaction between the two. In order to ascertain the effectiveness of STDAN, we conduct a substantial number of comparative experiments and ablation experiments on the Gaofen-2 data. Our method shows significant advantages in both qualitative and quantitative evaluations compared with state-of-the-art CCD algorithms. The main contributions of STDAN are as follows.

1) We propose a TDE module, which exploits the complementary advantages of pixelwise difference and channelwise concatenation operations. It can enhance the difference features while establishing the correlation between different temporal images, thereby alleviating the interference of task-independent changes, such as seasonal changes and illumination conditions.
2) CDFA is designed to continuously focus on the real changing area on the basis of enhanced difference features. It completes the aggregation between different levels of difference features in an incremental manner, and then further refines the difference features. Furthermore, the accurate CCD results can be obtained by integrating the different scale characteristics of CDEA through MRF.
3) Extensive experiments on Gaofen-2 cropland data confirm that the proposed method has strong performance in the detection of cropland changes. The remaining three mainstream datasets are used to conduct generalization experiments, proving that our method is equally applicable to other change detection scenarios.

II. Methodology

In this section, we furnish a comprehensive description of the proposed methodology. The motivation of STDAN is presented first. Subsequently, we outline the primary components of the model, including the TDE, CDFA, and MRF.

A. Motivation and Overview

In the scene of CCD, the crops in the cropland exhibit different characteristics in different seasons and light intensities. These pseudochanges result in numerous missed and false detections. Moreover, most changes in the cropland are small-scale man-made buildings, which account for a small proportion. That is to say, there exists a serious category imbalance between the changing building area and the nonchanging cropland area. Due to the influence of pseudochanges, these changed building areas may exhibit identical characteristics as the nonchanged cropland area, which further increases the difficulty of detection. In light of the aforementioned factors, we propose STDAN, which aims to alleviate the interference of pseudochanges to focus on the real changed areas and then detect the changed areas in the cropland with precision.

The flowchart for the proposed STDAN is shown in Fig. 2. The STDAN provides a straightforward and efficient solution to the issue of pseudochange and class imbalance in the CCD task by focusing on the genuine change area subsequent to the extraction of the shallow bitemporal features. In particular, taking into account the difference between bitemporal images, we initially employ Siamese pretrained PVTv2 [33] as encoders to extract the shallow features of different temporal images in parallel. The encoder consists of four stages, each including a patch embedding layer and a multihead attention mechanism. Following a pyramid structure, the output resolution of the four stages progressively shrinks from high (4 strides) to low (32 strides). It is noteworthy to mention that, during the subsequent procedure, we utilize the features generated by the first three stages of the encoder, and their spatial scales are 64×64, 32×32, and 16×16. In contrast to other methods that continuously extract deep features, we focus on the change area based on the above. Specifically, TDE is employed to enhance the change region of

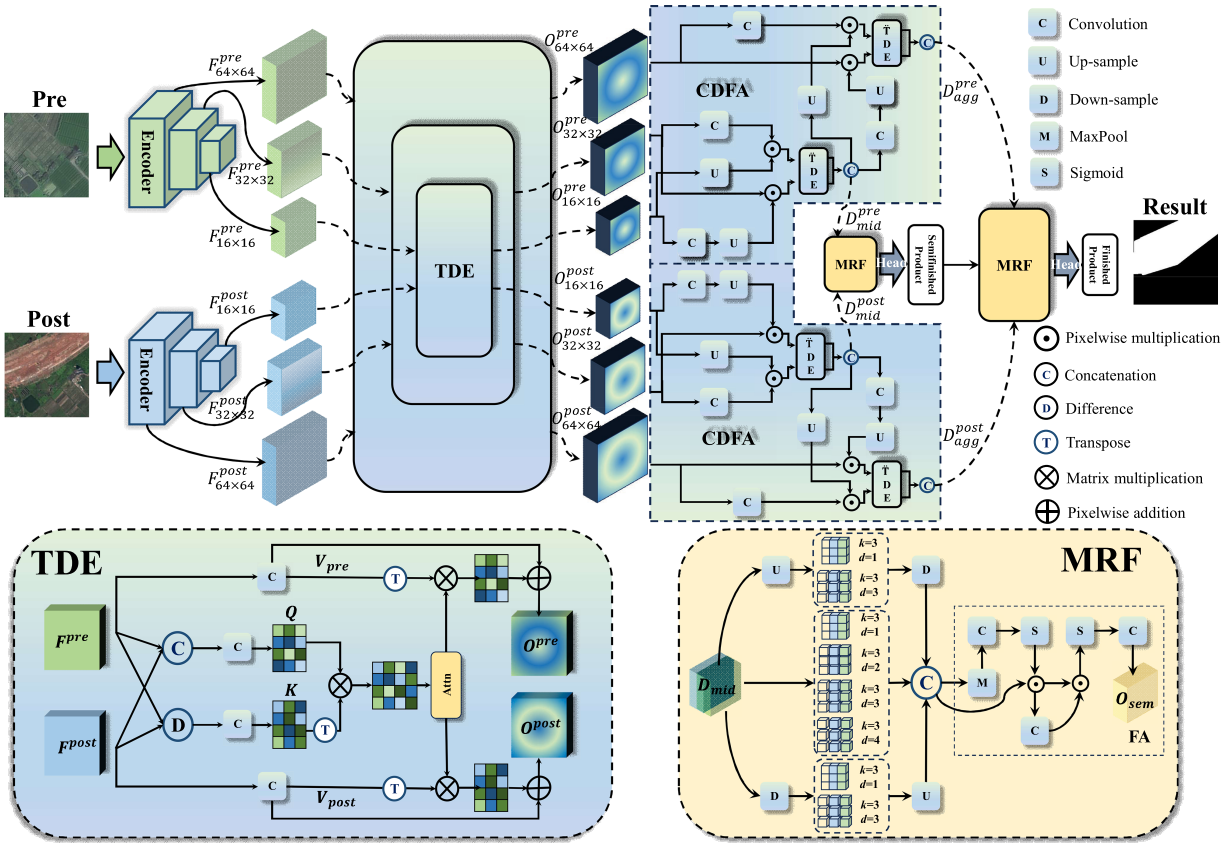

Fig. 2. Flowchart of the proposed STDAN.

the extracted features of the same dimension. In this process, we also consider the correlation between features of the same scale at the same time to avoid the interference of pseudochanges. Subsequently, the enhanced multilevel difference features are fed into the CDFA, which further focuses on the real changing regions in an incremental learning manner. Finally, the accurate CCD results can be obtained by incorporating the different scale characteristics of CDEA via MRF. In general, our proposed STDAN focuses on the real change area from beginning to end, which is very important to alleviate the interference of pseudochange and class imbalance problems and achieve the accurate CCD outcomes.

B. Cross-Temporal Difference Feature Enhancement

Most existing methods use the designed module to further extract the deep semantic features after extracting the shallow features of the bitemporal image by the encoders. However, owing to the influence of category imbalance, the changing part of the cropland area often accounts for only a small portion. This approach can result in a large number of redundant features, rendering it challenging to accurately detect the real changing area. To this end, after extracting the shallow features of the bitemporal image, we directly focus on the change area of different temporal features. Specifically, a simple difference operation, namely pixelwise subtraction, is used to capture the differences between bitemporal features. The above process can be expressed using the concept of set

$$D_{s\times s} = \complement \cup \left(F^{\text{pre}}_{s\times s} \cap F^{\text{post}}_{s\times s}\right), \text{s.t. } s \in \{64, 32, 16\} \quad (1)$$

where $D_{s\times s}$ represents the result of pixelwise subtraction of different temporal features with spatial scale s. $\complement$, $\cup$, and $\cap$ denote the complement, union, and intersection operations, respectively. F^{pre} and F^{post} represent the different temporal features. Considering the interference of pseudochanges, such as seasonal changes and light intensity in the CCD scene, simple difference operation cannot extract the real change area. To reduce the interference of task-irrelevant factors, it is essential to establish a correlation between different temporal difference features while capturing differences. To this end, we guide the model to enhance the real changed cropland area by simultaneously establishing the correlation and difference between the bitemporal images. Specifically, the channelwise concatenation operation is adopted to establish the correlation between bitemporal features by the learning manner. The above concatenation operation can be expressed as follows:

$$C_{s\times s} = F^{\text{pre}}_{s\times s} \cup F^{\text{post}}_{s\times s} \quad (2)$$

where $C_{s\times s}$ represents the result of channelwise concatenation of bitemporal features with a spatial scale s. Subsequently, we use the concept of cross attention [34] to establish the dependency relationship between the difference features and the global features of different temporals so as to enhance the captured

difference features while establishing the global correlation. The feature is first mapped through matrix multiplication and reshaped to query *Q*, key *K*, and value *V*. The above process can be expressed as follows:

$$\begin{aligned} Q &= P_Q \cdot D \\ K &= P_K \cdot C \\ V_{\text{pre}} &= P_V^{\text{pre}} \cdot F^{\text{pre}} \\ V_{\text{post}} &= P_V^{\text{post}} \cdot F^{\text{post}} \end{aligned} \tag{3}$$

where P_Q, P_K, and P_V represent the projection matrix, which is implemented by a convolution layer with kernel of 1. The similarity between global-difference features for both bitemporal features is allocated using two parallel branches, which is different from ordinary cross attention. The above process can be expressed as follows:

$$\begin{aligned} \text{Attn} &= \text{Softmax}\left(QK^{\text{T}}\right) \\ O^{\text{pre}} &= \text{Attn} \cdot V_{\text{pre}}^{\text{T}} + F^{\text{pre}} \\ O^{\text{post}} &= \text{Attn} \cdot V_{\text{post}}^{\text{T}} + F^{\text{post}} \end{aligned} \tag{4}$$

where Softmax and *T* represent the Softmax activation function and transpose operation, respectively. In general, the acquisition of global-difference features across the bitemporal can be accomplished through the guidance provided by the proposed TDE. This is conducive to alleviating task-independent pseudochange interference, thereby enhancing the difference characteristics.

C. Cross-Level Difference Feature Aggregation

In TDE, difference features in bitemporal images are enhanced. In order to further focus on the genuine changing region, we design CDFA, which aims to align different levels of changing feature representations within the same temporal feature through incremental learning. Initially, we continue to execute difference and concatenation operations on the same level of features to obtain difference features while simultaneously establishing correlation. Different levels of features contain rich semantic features that aid in noise suppression and focus on changing regions. In light of the fact that the pixel value of the changing region is larger than that of the nonchanging region, the pixelwise multiplication is adopted to further highlight the changing regions at different levels. At the same time, this approach can also further suppress the task-independent nonchanging regions (regions with lower pixel values). Fig. 2 shows the process of incremental aggregation of multilevel difference features. Taking difference features $O_{s_1}^{\text{pre}}$ and $O_{s_2}^{\text{pre}}$ as examples, the incremental aggregation of difference features can be expressed as follows:

$$D_{\text{mid}}^{\text{pre}} = \ddot{\text{TDE}} \begin{pmatrix} \text{UP}\left(\text{Conv}_{k=1}\left(O_{s_1}^{\text{pre}}\right)\right) \odot O_{s_2}^{\text{pre}} \\ \text{UP}\left(O_{s_1}^{\text{pre}}\right) \odot \text{Conv}_{k=1}\left(O_{s_2}^{\text{pre}}\right) \end{pmatrix} \tag{5}$$

where $\text{Conv}_{k=1}$, UP and $\odot$ represent the convolution operation with kernel of 1, upsampling operation, and pixelwise multiplication operation, respectively. In addition, s_1 and s_2 represent the spatial scales, which are 16×16 and 32×32, respectively. After incrementally aggregating different levels of features, $\ddot{\text{TDE}}$ is adopted to further focus on the changing regions while establishing global dependencies. In order to perform incremental learning more effectively, we replace the difference and concatenation operations in TDE with addition and multiplication operations. On this basis, $D_{\text{mid}}^{\text{pre}}$ and low-level feature $O_{s_3}^{\text{pre}}$ perform difference incremental aggregation again, and the process can be expressed as follows:

$$D_{\text{agg}}^{\text{pre}} = \ddot{\text{TDE}} \begin{pmatrix} \text{UP}\left(\text{Conv}_{k=1}\left(D_{\text{mid}}^{\text{pre}}\right)\right) \odot O_{s_3}^{\text{pre}} \\ \text{UP}\left(D_{\text{mid}}^{\text{pre}}\right) \odot \text{Conv}_{k=1}\left(O_{s_3}^{\text{pre}}\right) \end{pmatrix} \tag{6}$$

where $D_{\text{agg}}^{\text{pre}}$ represents the result of difference incremental aggregation of three different levels of features. s_3 represents the spatial scale, and its size is 64×64. Through feature aggregation step by step, the difference features of various scales in the same temporal image are comprehensively considered, and the real change area is then accurately located.

D. MRF and Constraint

Following the CDFA, the bitemporal difference features of two different scales, namely D_{mid} and D_{agg}, are obtained. Considering the difference features of different scales, we adopt atrous convolutions with different dilation rates. The size of the receptive field can be expanded without increasing the parameters with atrous convolution. Specifically, in order to adapt to the characteristics of different scales, we use four parallel branches with different dilation rates to extract the difference features of different scales in the different temporal features. Taking bitemporal difference features $D_{\text{mid}}^{\text{pre}}$ and $D_{\text{mid}}^{\text{post}}$ as an example, the process of MRF can be expressed as follows:

$$D_{\text{mid}} = D_{\text{mid}}^{\text{pre}} \cup D_{\text{mid}}^{\text{post}} \tag{7}$$

$$M_i = \text{Conv}_{k=3,d=i}\left(D_{\text{mid}}\right), \text{ s.t. } i \in \{1,2,3,4\} \tag{8}$$

where d represents the dilation rate. To maximize the utilization of multiscale features, we simultaneously upsample and downsample D_{mid} and employ convolutions with different dilation rates to extract features and sample to the original scale. The process can be expressed as follows:

$$M_j = \text{Down}\left(\text{Conv}_{k=3,d=j}\left(\text{UP}\left(D_{\text{mid}}\right)\right)\right), \text{ s.t. } j \in \{1,3\} \tag{9}$$

$$M_k = \text{Up}\left(\text{Conv}_{k=3,d=k}\left(\text{Down}\left(D_{\text{mid}}\right)\right)\right), \text{ s.t. } k \in \{1,3\} \tag{10}$$

$$M = M_i \cup M_j \cup M_k. \tag{11}$$

By performing MRF on bitemporal difference features, a more detailed change representation can be obtained. Finally, we utilize the concept of incremental aggregation in CDFA to further highlight the areas of real change. The process can be expressed as follows:

$$O_{\text{mid}} = \sigma\left(\text{Conv}_{k=1}\left(\text{MP}\left(M\right)\right)\right) \odot M \tag{12}$$

$$O_{\text{sem}} = H\left(\text{Conv}_{k=1}\left(\sigma\left(\left(\text{Conv}_{k=1}\left(O_{\text{mid}}\right)\right) \odot O_{\text{mid}}\right)\right)\right) \tag{13}$$

where σ and MP represent the Sigmoid activation function and adaptive max pooling, respectively. Besides, H denotes the

classification head. In this way, we get the change detection results for the semifinished version. Subsequently, we execute (7)–(13) once more for $D_{\text{agg}}^{\text{post}}$ and $D_{\text{agg}}^{\text{post}}$, resulting in the final result O_{final}.

Correspondingly, we simultaneously constrain the semifinished product O_{sem} and the finished product O_{final}. Specifically, we use cross entropy as the loss function to constrain the output of the model

$$L_{\text{sem}} = -\frac{1}{N}\sum_{n=1}^{N}\left[G\log O_{\text{sem}} + (1-G)\log\left(1-O_{\text{sem}}\right)\right] \tag{14}$$

$$L_{\text{final}} = -\frac{1}{N}\sum_{n=1}^{N}\left[G\log O_{\text{final}} + (1-G)\log\left(1-O_{\text{final}}\right)\right] \tag{15}$$

$$L_{\text{Total}} = \alpha_1 L_{\text{sem}} + L_{\text{final}} \tag{16}$$

where N represents the total number of pixels in the cropland area of the dataset. G denotes the ground truth. The hyperparameter α_1 is used to adjust the ratio between O_{sem} and O_{final}, which we will discuss in Section III-F.

E. Implementation Details and Evaluation Metrics

Taking into account the actual requirements of different application scenarios, we devise three different versions by using encoders with varying depths, designated STDAN-S, STDAN-M, and STDAN-L, respectively, each featuring a lower model complexity version, a medium model complexity version, and a higher model complexity version. Specifically, the encoders in STDAN-S, STDAN-M, and STDAN-L correspond to $b0$, $b1$, and $b2$ versions of PVTv2, respectively, where the number of channels in $b0$ is 32, 64, and 160, respectively, and the number of channels in $b1$ and $b2$ is 64, 128, and 320, respectively. In addition, in terms of parameter settings, STDAN can achieve convergence at 100 rounds under the premise of a batch size of 16. We use AdamW [35] as the optimizer and set the weight attenuation to 0.0004. The initial learning rate is 0.00025, and StepLR is used to update the learning rate every eight rounds to reduce it by half.

We select six mainstream metrics as the criteria for quantitative evaluation, which are precision (P), recall (R), $F1$-score ($F1$), intersection over union (IOU), overall accuracy (OA), and Kappa coefficient

$$P = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{17}$$

$$R = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{18}$$

$$F1 = \frac{2PR}{P+R} \tag{19}$$

$$\text{IOU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{20}$$

$$\text{OA} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{21}$$

$$P_e = \frac{(\text{TP}+\text{FP})(\text{TP}+\text{FN})+(\text{FP}+\text{TN})(\text{FN}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})^2} \tag{22}$$

$$\text{Kappa} = \frac{\text{OA}-P_e}{1-P_e} \tag{23}$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively. In addition, the units of these six metrics are all percentages, and the larger the better.

III. Experiments

A. Datasets

We use the public CCD dataset CL-CD [36] to verify the performance of the proposed STDAN. The dataset consists of 600 high-resolution satellite images of 512 × 512 sizes, which were collected by Gaofen-2 satellite in Guangdong Province, China, between 2017 and 2019, with a spatial resolution of 0.5–2 m. The CL-CD dataset has a variety of change detection types, including buildings, roads, lakes, and cropland. The original dataset contains 320 pairs for training, 120 pairs for validation, and 120 pairs for testing. In the experiment, we cut the data to the size of 256 × 256, and finally obtained 1280 pairs for training, 480 pairs for validation, and 480 pairs for testing.

B. Comparative Methods

To evaluate the efficacy of STDAN, we choose four mainstream categories comprising of 16 SOTA methods for comparison experiments, namely CNN-based methods (FC-EF [18], FC-Siam-Conc [18], FC-Siam-Diff [18], FresUNet [19], and CGNet [38]), CNN-attention-based methods (DSIFNet [20], SNUNet [21], DMINet [32], AANet [30], and B2CNet [39]), transformer-based approaches (BIT [23]), and CNN-transformer-based approaches (ICIFNet [37], WNet [28], and CSINet [31]). In addition, we have included two CCD baseline methods (MSCANet [43] and MeGNet [44]). For a fair comparison, we use the authors' officially released code. All of the codes are executed on a computer that is equipped with a GTX-3090 GPU. In addition, codes for STDAN will be published.[1]

C. Comparative Analysis

Table I lists the quantitative results of the CL-CD dataset. It can be seen that the scores obtained by our method on $F1$, IOU, OA, and Kappa are in a prominent position. It is worth mentioning that both the $F1$-score and the IOU are used as metrics to measure the overall performance of the model. Compared with other advanced methods, only the performance of AANet and MeGNet is close to our proposed STDAN. The three different versions of STDAN achieve 78.32%, 79.63%, and 79.42% on the $F1$-score, respectively. Compared with the suboptimal AANet and MeGNet, STDAN-M achieves 2.12% and 3.58% improvement, respectively. Compared with the suboptimal AANet on IOU, the three different versions of STDAN have achieved

[1][Online]. Available: https://github.com/bao11seven/STDAN-CD

TABLE I
QUANTITATIVE RESULTS ON THE CL-CD DATASET

Methods	Quantitative Analysis						Efficiency Analysis		
	P	R	F1	IOU	OA	Kappa	Params(M)	FLOPs(G)	Time(ms)
FC-EF$_{2018}$	55.13	63.60	59.06	41.91	93.44	55.52	**1.35**	**3.58**	**2.297**
FC-Siam-Diff$_{2018}$	62.56	57.89	60.13	42.99	94.29	57.06	**1.35**	**4.73**	**2.998**
FC-Siam-Conc$_{2018}$	58.39	61.99	60.14	43.00	93.88	56.83	1.55	5.33	3.063
FresUNet$_{2019}$	60.12	68.19	63.90	46.95	94.27	60.80	**1.10**	**2.02**	**2.414**
DSIFNet$_{2020}$	69.48	69.60	69.54	53.30	95.46	67.09	35.73	82.27	12.362
SNUNet$_{2022}$	68.75	68.04	68.39	51.97	95.32	56.87	12.03	54.83	11.141
BIT$_{2022}$	70.43	74.46	72.39	56.73	95.77	70.11	3.50	10.63	5.251
ICIFNet$_{2022}$	78.13	70.64	74.20	58.98	96.34	72.23	23.84	25.41	18.103
MSCANet$_{2022}$	71.96	75.34	73.61	58.24	95.98	71.44	16.80	14.80	6.223
DMINet$_{2023}$	76.17	75.92	76.05	61.35	96.44	74.12	6.24	14.55	4.668
CGNet$_{2023}$	70.16	70.29	70.23	54.12	95.57	67.83	33.68	82.23	9.194
WNet$_{2023}$	69.10	**83.17**	75.48	60.62	95.98	73.32	43.07	19.20	17.892
AANet$_{2024}$	**81.93**	73.55	77.51	63.28	96.82	75.81	7.91	24.22	6.354
B2CNet$_{2024}$	72.28	75.28	73.75	58.41	96.01	71.59	16.10	22.36	6.590
CSINet$_{2024}$	68.76	72.78	70.71	54.70	95.51	68.29	62.18	367.25	33.496
MeGNet$_{2024}$	80.09	73.07	76.42	61.83	96.64	74.61	41.07	22.52	19.472
STDAN-S	81.52	**75.35**	**78.32**	**64.36**	**96.90**	**76.65**	3.11	4.88	7.951
STDAN-M	**81.82**	**77.56**	**79.63**	**66.16**	**97.05**	**78.04**	12.31	16.02	8.887
STDAN-L	**82.64**	76.44	**79.42**	**65.86**	**97.05**	**77.83**	20.49	20.47	13.122

Red, green and blue highlights indicate the best, the second-best, and the third-best values, respectively.

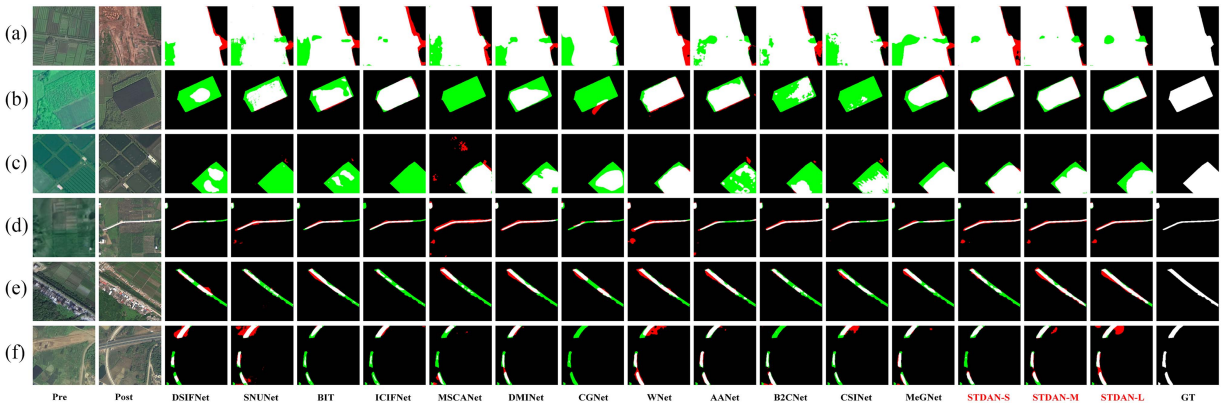

Fig. 3. Qualitative experimental results on the CL-CD dataset.

1.08%, 2.88%, and 2.58% improvement, respectively. In terms of OA, STDANe-S and STDANet-M are slightly weaker than AANet, our method still achieves great improvement compared with other methods. Specifically, STDAN achieves an increase of 3.39% and 3.69%, respectively, compared with ICIFNet. In addition, STDAN-L achieved the best OA of 82.64%. This demonstrates that our method improves the accuracy of cropland change by focusing more on the identification of change characteristics. Our method's accuracy gradually improves as the network's depth improves, indicating that the enlarging of the model's scope also aids in further alleviating the impact of pseudochanges, thereby enhancing the proposed module's focusing on the identifying real change areas. In addition, the deepening of the number of network layers will also bring a slight decline in performance. For example, the score of STDAN-L in R is lower than that of STDAN-M, resulting in a decrease on $F1$ and IOU. This may be attributable to the more extensive redundancy features that disturb the model's attention to the real change area. In general, our method can achieve excellent detection results on the CL-CD dataset.

Fig. 3 depicts the qualitative experimental results, wherein red indicates the misdetection and green indicates the missed detection. A variety of CCD scenarios are displayed, such as land development, water body change, and vegetation growth. As shown in Fig. 3(a), vegetation characterization shows a large difference due to seasonal changes and light intensity, which results in missed detection by most methods. Only WNet and the proposed STDAN do not miss detection. In contrast to WNet,

STDAN exhibits greater accuracy in predicting the boundary of the change area and exhibits no false detection, indicating that our method possesses superior change feature focusing ability while guaranteeing detection accuracy. Fig. 3(b) and (c) shows the transformation of cropland into a water body, necessitating an enhanced analysis of the nonlinear change and reflection characteristics of the water body's boundary by the model. It is evident that most methods have issues with detection and are more likely to lose the change boundary. Our method, on the other hand, performs more stably in identifying such scenes, which is similar to GT, thanks to our method's ability to focus on real changing areas. Fig. 3(d) and (e) shows the urban expansion and road construction scenarios resulting from the development of urbanization. The shape of this type of scene change area is irregular, and the change sample only accounts for a small portion of the image, rendering it challenging to detect accurately. Our STDAN performs more stable and can predict the change area more accurately and completely. In addition, there is a problem of inconsistent resolution between the bitemporal images in Fig. 3(d), which leads to error detection in most methods. It can be observed that only B2CNet and STDAN are more accurate and complete for road detection, and there are no significant numbers of incorrect detections. Furthermore, our method detects errors in the lower left corner region. By comparing bitemporal images in Fig. 3(d), it can be observed that the lower left corner region exhibits a change; however, it is not marked on the label. It is evident that STDAN possesses the ability to actively concentrate on the genuine change domain. To summarize, our method has stronger dynamic monitoring capabilities in the CCD scene and can achieve better detection outcomes.

D. Efficiency Analysis

Performance and efficiency are important for practical applications. The proposed STDAN has achieved good results in detection performance. In this section, we continue to explore the efficiency of each method and the relationship between performance and efficiency. As shown in Table I, we test the efficiency of each method from three aspects, including the number of model parameters (Params), the number of floating points (FLOPs), and the test time (Time). In terms of Params, the proposed STDAN-S has significant advantages, second only to the four methods FC-EF, FC-Siam-Diff, FC-Siam-Conc, and FresUNet. As the network depth increases, the Params of STDAN-M and STDAN-L increase, but they still have advantages over DSIFNet, ICIFNet, CGNet, WNet, and B2CNet. In terms of FLOPs, STDAN-S is second only to FC-EF, FC-Siam-Diff, and FresUNet. Even if STDAN-M and STDAN-L have increased in FLOPs compared with STDAN-S, they still have advantages over other methods. In terms of Time, the proposed method shows disadvantages, only better than DSIFNet, SNUNet, ICIFNet, WNet, and CSINet. In practice, a lower number of parameters and computation is a prerequisite for performing detection tasks in resource-constrained hardware environments. Lightweight processing, such as model pruning and distillation, can be performed on the model. In addition, images can be segmented and then a processing pipeline can be constructed to minimize the processing time in practical applications. In addition, as shown in Fig. 4, we analyze the relationship between detection performance and efficiency, where the area of each node in Fig. 4(a) and (b) represents the parameters and FLOPs, respectively. It can be seen from Fig. 4(a) that STDAN-M achieves the best balance among the three. Similar results can be seen in Fig. 4(b), STDAN-M also achieves the best balance among IOU, Time, and Params.

E. Generalization Analysis

The proposed STDAN shows high performance in both quantitative and qualitative evaluation of CCD scenarios. In order to verify the generalization ability of STDAN in other change detection scenarios, we conduct corresponding experiments on three mainstream datasets, including WHU-CD [40], LEVIR-CD [41], and SYSU-CD [42]. The WHU-CD includes a pair of 32 507 $\times$ 15 354 high-resolution aerial images with a spatial resolution of 0.2 m. The data mainly include building damage changes in Christchurch, New Zealand, after the 2012 earthquake. In our experiment, we cut the image into 256 $\times$ 256 sizes and divided the data into 6096 image pairs for training, 762 image pairs for validation, and 762 image pairs for testing at a ratio of 8:1:1. LEVIR-CD includes 637 pairs of high-resolution satellite images with a size of 1024 $\times$ 1024, and its spatial resolution is 0.5 m. The study area mainly covers the architectural and land cover changes in 20 different urban areas in Texas from 2002 to 2018. We divide the image into image blocks with a spatial scale of 256 $\times$ 256. Finally, 7120 image pairs are obtained for training, 1024 images for validation set, and 2048 images for testing. SYSU-CD consists of 20 000 pairs of high-resolution aerial images of 256 $\times$ 256 scales, with a spatial resolution of 0.5 m. The dataset covers various types of changes in Hong Kong, such as ports, suburbs, and high-rise buildings. In the experiment, we divide the data into 12 000 image pairs for training, 4000 image pairs for validation, and 4000 image pairs for testing at a ratio of 6:2:2.

Quantitative experimental results are shown in Table II. On the WHU-CD dataset, STDAN achieves optimal and suboptimal scores on all metrics. Compared with other methods, our method achieves at least 0.47%, 1.39%, 1.35%, and 2.53% higher scores on P, R, $F1$, and IOU. On the LEVIR-CD dataset, STDAN achieves the highest score on $F1$, although it was slightly lower than CGNet on P and slightly lower than DSIFNet on R. Specifically, compared with other methods, our method achieves at least 0.39% and 0.74% higher scores on $F1$ and IOU. On the SYSU-CD dataset, except for R, STDAN achieves optimal and suboptimal scores on other metrics. Compared with other methods, our method achieves at least 1.53%, 1.30%, and 2.24% higher scores on P, $F1$, and IOU. These experimental results demonstrate that STDAN also has better detection performance in mixed building areas, owing to focusing on the real change area while establishing the correlation between different temporal features. Figs. 5–7 show the corresponding qualitative experimental results. It is evident that the proposed STDAN shows the lowest missed detection and false detection rates in the

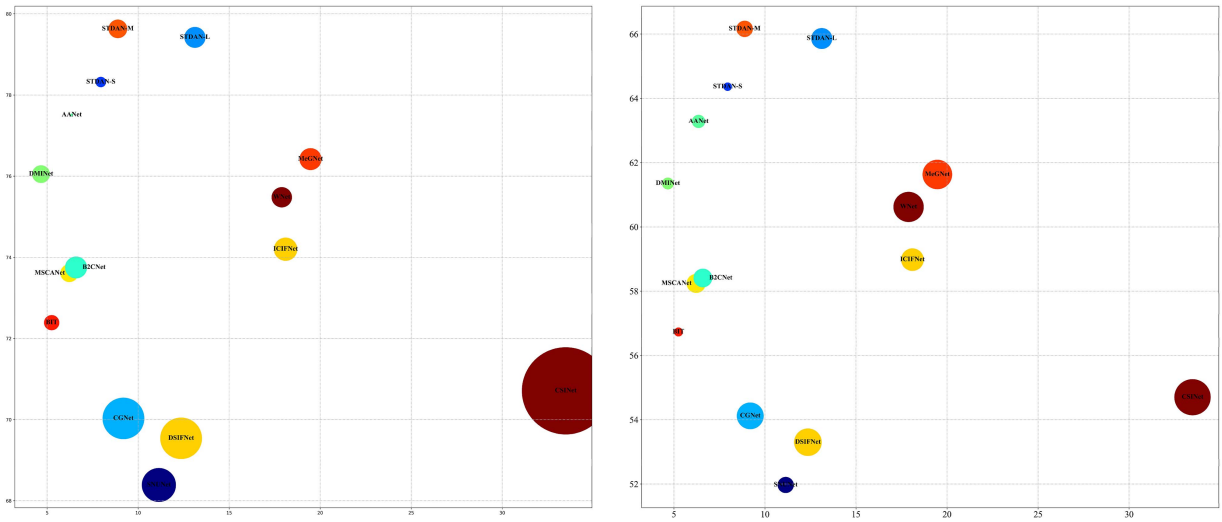

Fig. 4. Comparison of the efficiency of existing state-of-the-art DL-based CCD methods.

TABLE II
QUANTITATIVE RESULTS ON THREE MAINSTREAM DATASETS

Methods	WHU-CD						LEVIR-CD						SYSU-CD					
	P	R	F1	IOU	OA	Kappa	P	R	F1	IOU	OA	Kappa	P	R	F1	IOU	OA	Kappa
FC-EF$_{2018}$	67.37	78.76	72.62	57.01	97.72	71.44	78.31	73.56	75.86	61.11	97.61	74.61	72.50	78.25	75.27	60.34	87.87	67.25
FC-Siam-Diff$_{2018}$	74.54	76.19	75.35	60.45	98.08	74.36	83.87	76.64	80.09	66.80	98.06	79.07	85.76	58.32	69.42	53.17	87.89	62.21
FC-Siam-Conc$_{2018}$	73.93	76.84	75.36	60.46	98.07	74.35	82.65	79.72	81.16	68.29	98.11	80.17	72.24	81.42	76.55	62.01	88.24	68.74
FresUNet$_{2019}$	83.36	78.45	80.83	67.83	98.57	80.09	87.54	85.89	86.71	76.54	98.66	86.00	75.84	**84.07**	79.74	66.31	89.93	73.06
DSIFNet$_{2020}$	91.45	85.29	88.26	78.99	99.13	87.81	89.93	**90.89**	90.41	82.50	99.02	89.89	82.81	73.87	78.08	64.05	90.22	71.82
SNUNet$_{2022}$	88.86	87.67	88.26	78.99	99.10	87.80	91.34	88.63	89.96	81.76	98.99	89.43	77.32	80.11	78.69	64.87	89.77	71.96
BIT$_{2022}$	92.96	88.04	90.43	82.53	99.28	90.06	91.51	88.05	89.75	81.40	98.98	89.21	79.86	76.99	78.40	64.47	89.99	71.89
ICIFNet$_{2022}$	**94.90**	87.63	91.12	83.69	99.34	90.78	91.41	89.02	90.20	82.15	99.01	89.68	78.44	78.40	78.42	64.50	89.82	71.76
MSCANet$_{2022}$	91.75	89.27	90.49	82.63	99.28	90.12	91.17	87.29	89.19	80.48	98.92	80.48	78.74	75.30	76.98	62.57	89.38	70.08
DMINet$_{2023}$	95.61	88.76	92.06	85.28	99.41	91.75	92.12	89.29	90.68	82.95	99.07	90.19	82.24	81.47	81.85	69.28	91.48	76.29
CGNet$_{2023}$	93.65	89.16	91.35	84.08	99.35	91.02	**92.95**	90.12	**91.51**	**84.35**	**99.15**	**91.06**	80.26	79.18	79.72	66.28	90.50	73.52
WNet$_{2023}$	92.50	**90.87**	91.68	84.64	99.37	91.35	88.72	90.75	89.72	81.36	98.94	89.17	75.70	**85.28**	80.20	66.95	90.07	73.61
AANet$_{2024}$	93.23	89.29	91.21	83.85	99.34	90.87	91.19	89.24	90.21	82.16	99.01	89.69	82.16	81.96	82.06	69.58	91.55	76.53
B2CNet$_{2024}$	93.70	90.97	**92.31**	**85.73**	**99.42**	**92.01**	91.27	**91.01**	91.14	83.72	99.10	90.66	81.80	**84.71**	83.23	71.28	91.95	77.94
CSINet$_{2024}$	93.55	89.78	91.63	84.55	99.37	91.30	91.97	90.50	91.23	83.88	99.11	90.77	78.79	81.09	79.93	66.57	90.39	73.62
MeGNet$_{2024}$	93.51	90.41	91.93	85.07	99.39	91.61	91.04	90.28	90.66	82.91	99.05	90.16	84.54	75.39	79.70	66.25	90.94	73.90
STDAN-S	93.82	90.59	92.17	85.48	99.41	91.87	92.42	89.93	91.16	83.76	99.11	90.69	**84.92**	81.87	**83.37**	**71.48**	**92.30**	**78.36**
STDAN-M	**94.55**	**92.23**	**93.37**	**87.57**	**99.50**	**93.11**	**92.77**	90.82	**91.78**	**84.81**	**99.17**	**91.32**	**85.93**	82.35	**84.10**	**72.56**	**92.66**	**79.33**
STDAN-L	**96.06**	**91.19**	**93.56**	**87.90**	**99.52**	**93.31**	**92.89**	**90.88**	**91.87**	**84.97**	**99.18**	**91.44**	**87.07**	81.72	**84.31**	**72.88**	**92.83**	**79.67**

Red, green and blue highlights indicate the best, the second-best, and the third-best values, respectively.

three datasets, which further proves that STDAN is also suitable for mixed building change detection scenarios.

F. Ablation Analysis

In order to evaluate the effectiveness of the proposed TDE, CDFA, and MRF, we conduct ablation experiments on four datasets, including LEVIR-CD, WHU-CD, SYSU-CD, and CL-CD. All variations in the ablation experiments used PVTv2 as the encoder, and the variants of our proposed methods are used as experimental comparisons. The baseline method contains only TDE and uses feature difference as the decoder after passing through the TDE module. Table III lists the quantitative results of ablation experiments. It can be seen that after adding the CDFA module to the baseline method, the $F1$-score has improved by 0.20%, 0.16%, 0.11%, and 0.73% on the four datasets. For variant 2, compared with STDAN removing the CDFA module, the $F1$-score decreased by 0.40%,0.31%, 1.18%,

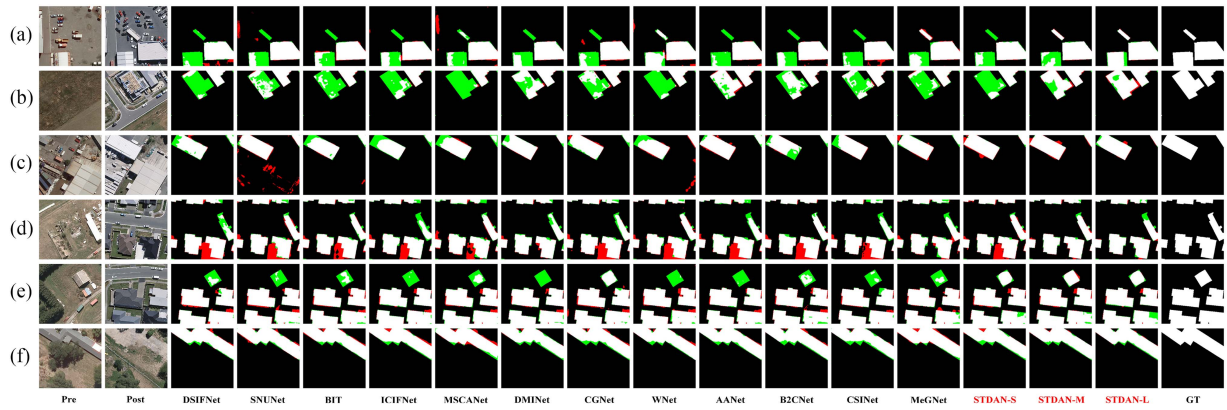

Fig. 5. Qualitative experimental results on the WHU-CD dataset.

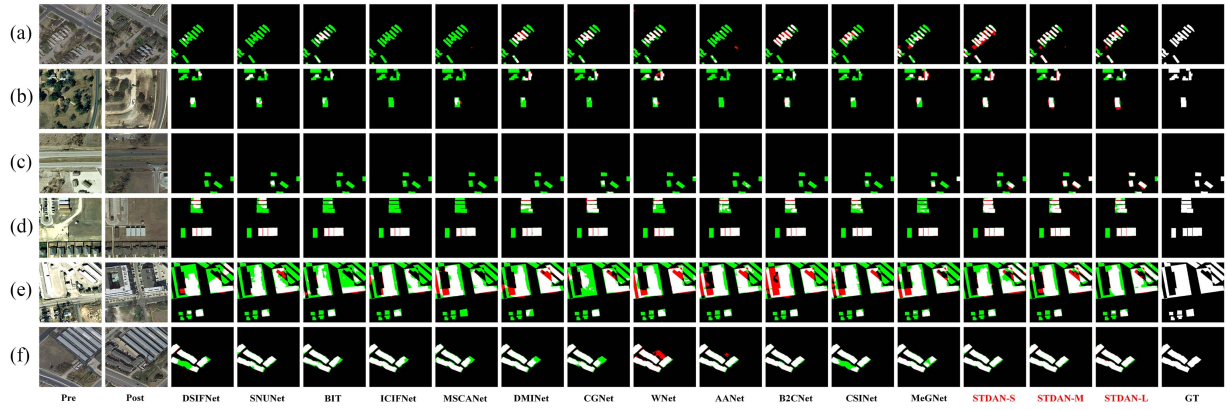

Fig. 6. Qualitative experimental results on the LEVIR dataset.

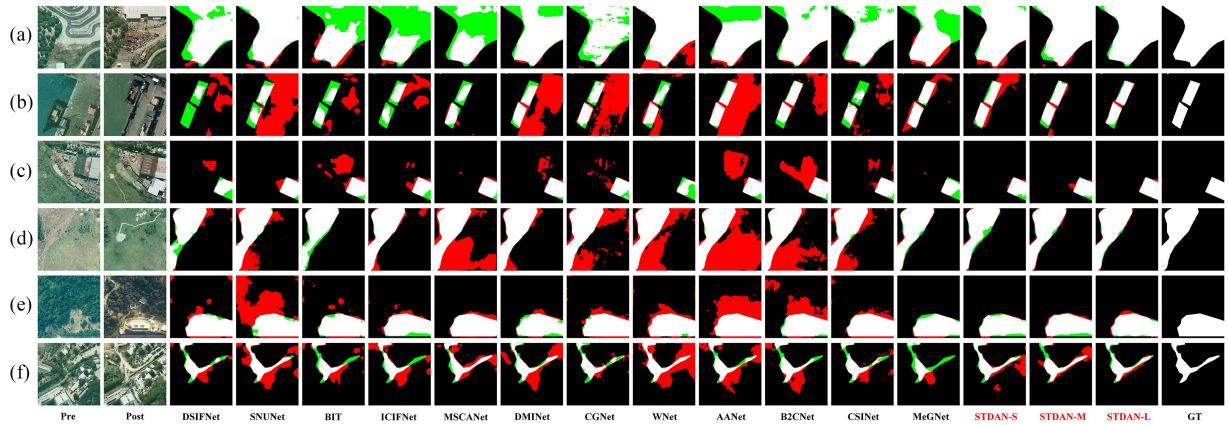

Fig. 7. Qualitative experimental results on the SYSU dataset.

and 0.82% on the four datasets, respectively. Especially on the CLCD and SYSU-CD datasets, the impact of the CDFA module is more obvious. This shows that the CDFA module helps to suppress the influence of noise and pays more attention to the change area by combining different levels of semantic features. Due to the more types of CL-CD scene changes and the more complex scenes, the pseudochanges have a stronger impact on the experiment. By comparing variant 3 with STDAN, it can be seen that, after removing the TDE module, our method reduces the IOU of the four datasets by 0.73%, 0.99%, 0.68%, and 1.60%, respectively. This shows the effectiveness of the TDE module, which can enhance the difference features between images at different temporal features, establish the correlation between images at different temporal, and suppress the influence of pseudochanges. By comparing variant 1 with STDAN, after adding the MRF module, the four datasets have achieved

TABLE III
ABLATION EXPERIMENTAL RESULTS OF EACH MODULE ON FOUR DATASETS

Ablation				Datasets							
				LEVIR-CD		WHU-CD		SYSU-CD		CLCD	
Variantion	TDE	CDFA	MRF	F1	IOU	F1	IOU	F1	IOU	F1	IOU
Baseline	√			91.29	83.97	92.73	86.45	82.50	70.21	78.01	63.94
Variant I	√	√		91.49	84.31	92.89	86.73	82.61	70.37	78.73	64.92
Variant II	√		√	91.38	84.13	93.06	87.03	82.92	70.82	78.81	65.04
Variant III		√	√	91.35	84.08	92.81	86.58	83.64	71.88	78.46	64.56
STDAN	√	√	√	**91.78**	**84.81**	**93.37**	**87.57**	**84.10**	**72.56**	**79.63**	**66.16**

Red highlights indicate the best values.

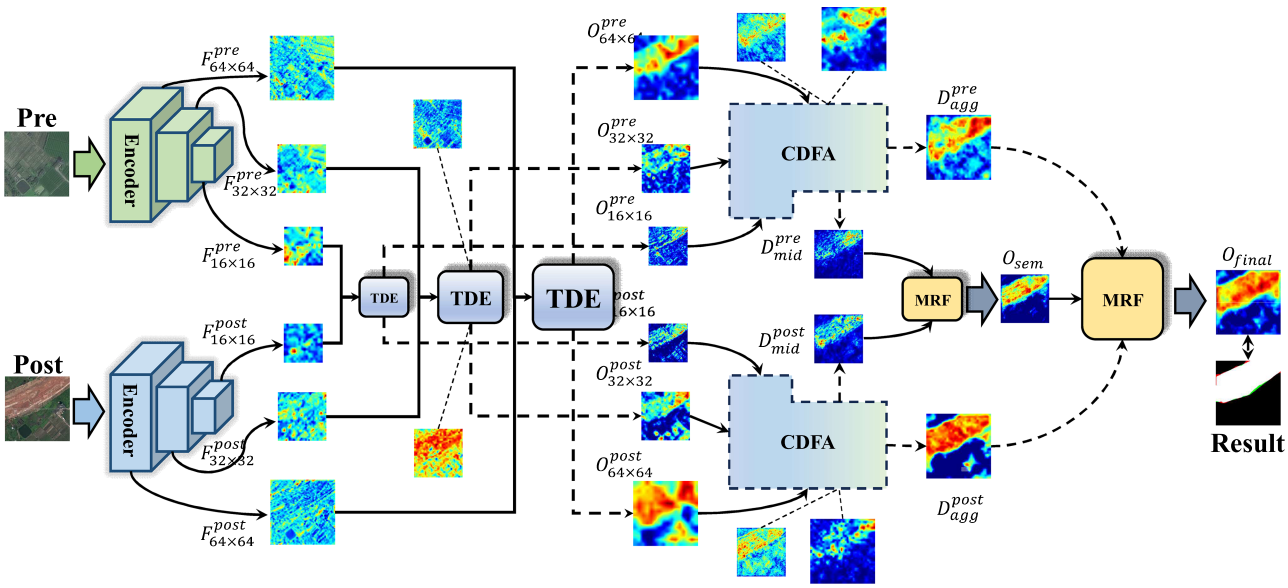

Fig. 8. Visualization of intermediate features in the proposed STDAN.

considerable improvement in $F1$ and IOU. The reason for this is that MRF is better adapted to the changing regions of different scales and provides more suitable receptive fields while fitting the changing regions, leading to more accurate change detection results. As a result, extensive experiments demonstrate the effectiveness of each module in STDAN, especially when confronted with CCD tasks.

In order to investigate the impact of deep supervision on model training, we conduct corresponding experiments on the constraint function in (16). The experimental results are shown in Table IV. As the value of α_1 increases, it can be seen that the proposed STDAN has achieved performance improvement on both LEVIR-CD and WHU-CD datasets, but on SYSU-CD and CLCD, it can be observed that it has decreased somewhat. This may be attributable to excessive supervision, resulting in the loss of detailed features, thereby reducing the model's performance. STDAN-M achieves the optimal performance on LEVIR-CD, SYSU-CD, and CLCD datasets when the value of α_1 is 0.5. In general, the implementation of the deep supervision is an effective way to improve the performance of the model. We have determined the most suitable parameter design through experiments. In the aforementioned experiments, we set the value of α_1 to 0.5 to attain considerable performance in different detection scenarios.

TABLE IV
ABLATION EXPERIMENTAL RESULTS OF LOSS FUNCTIONS ON FOUR DATASETS

α_1	LEVIR-CD		WHU-CD		SYSU-CD		CLCD	
	F1	IOU	F1	IOU	F1	IOU	F1	IOU
0	91.60	84.49	92.47	85.99	83.22	71.26	78.26	64.28
0.2	91.65	84.58	**93.49**	**87.78**	83.34	71.44	79.21	65.58
0.5	**91.78**	**84.81**	93.37	87.57	**84.10**	**72.56**	**79.63**	**66.16**
0.8	91.67	84.63	93.35	87.53	84.00	72.42	78.75	64.94
1.0	91.65	84.58	92.93	86.79	83.04	71.00	78.23	64.24

Red highlights indicate the best values.

IV. DISCUSSION

In the scenario of CCD, crops are greatly affected by seasonal changes and light density. Moreover, there is a serious class imbalance between the changing building area and the nonchanging cropland area, and the two may exhibit identical representation under different light conditions. Therefore, it is imperative for the model to exclude the interference of pseudochanges and concentrate on the real changing area. The

existing DL-based CCD methods continue to extract deep semantic features based on the features extracted by the encoder. The mode's inability to focus on the actual changing areas is hampered by these redundant features, which in turn results in its high model complexity. In order to accurately detect the real change area in cropland, we propose the STDAN for CCD. Initially, STDAN enhances the difference features while establishing the correlation between different temporal features using the TDE, which can suppress task-independent interference. Subsequently, the CDFA enables the aggregation between different levels of difference features in an incremental manner, thereby further refining the change area. Ultimately, the accurate results can be obtained by integrating the difference characteristics at different scales in CDFA through MRF. The results of Section III-C demonstrate that the proposed STDAN is capable of achieving more accurate detection outcomes in CCD scenes. Section III-D validates the efficacy of STDAN. In Section III-E, we analyze the generalization of the proposed STDAN on three mainstream mixed building datasets. The experimental results show that STDAN is also suitable for building change detection scenarios. Section III-F verifies the effectiveness of each module in STDAN. In addition, we visualize the intermediate features of the model. An instance with images from the CL-CD dataset is delivered, as shown in Fig. 8. It can be seen that the shallow features extracted by the encoder are coarse and it is difficult to distinguish between changing and nonchanging regions. Through the guidance of TDE, enhanced difference features are obtained and the outline of the change region can be roughly seen. Subsequently, using CDFA, different scale features of the same temporal are aggregated, revealing the first glimpse of the change region. Finally, the accurate results can be obtained by integrating the different characteristics at different scales in CDFA through MRF. It is evident that the proposed STDAN can focus on the real changing area, thereby achieving the accurate detection results.

V. Conclusion

In view of the difference in crop characterization caused by pseudochange and the similarity between the change area and the nonchange area caused by the imbalance of positive and negative samples in the CCD scene, we propose STDAN, which can alleviate the interference of pseudochange to focus on the real change area, so as to achieve the accurate detection results. Through a multitude of comparative and efficiency analyses conducted on the cropland data collected by the Gaofen-2 sensor, it has been proved that STDAN can achieve the accurate CCD results while maintaining high efficiency. In addition, we conduct generalization experiments on three mainstream mixed building datasets. The corresponding quantitative and qualitative evaluation results prove that STDAN is also suitable for building change detection scenarios. Finally, we conduct ablation experiments on the above four datasets to verify the effectiveness of each module. However, in terms of testing time, the proposed method still exhibits significant room for improvement. In addition, this work primarily focuses on the binary CCD task, which is incapable of providing specific information on the type of changes in cropland. Therefore, it is necessary to further explore the semantic CCD of different vegetation in cropland under different environmental conditions. In future work, we will further improve the fusion performance and efficacy. Furthermore, we will further explore the linkage of high-resolution series data in different tasks to further promote the development of practical applications. In particular, we will exploit the advantageous information of the multimodal data provided by the Gaofen series sensors, such as the panchromatic and multispectral imagery provided by Gaofen-2 and the synthetic aperture radar imagery provided by Gaofen-3, and then carry out multimodal fusion and change detection.

Declaration of Competing Interest

There are no relevant financial or nonfinancial competing interests to report.

Acknowledgment

The authors would like to thank the anonymous reviewers and members of the editorial team for their comments and suggestions.

References

[1] D. Lu, Z. Wang, K. Su, Y. Zhou, X. Li, and A. Lin, "Understanding the impact of cultivated land-use changes on China's grain production potential and policy implications: A perspective of non-agriculturalization, non-grainization, and marginalization," *J. Cleaner Prod.*, vol. 436, Jan. 2024, Art. no. 140647.

[2] W. Chen, L. Yang, J. Zeng, J. Yuan, T. Gu, and Z. Liu, "Untangling the increasing elevation of cropland in China from 1980 to 2020," *Geogr. Sustain.*, vol. 4, no. 4, pp. 281–293, Dec. 2023.

[3] B. Xu et al., "Exploring the potential of Gaofen-1/6 for crop monitoring: Generating daily decametric-resolution leaf area index time series," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 4401614.

[4] W. Chen and G. Liu, "A novel method for identifying crops in parcels constrained by environmental factors through the integration of a Gaofen-2 high-resolution remote sensing image and Sentinel-2 time series," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 450–463, Jan. 2024.

[5] G. Cheng et al., "Change detection methods for remote sensing in the last decade: A comprehensive review," *Remote Sens.*, vol. 16, no. 13, Jun. 2024, Art. no. 2355.

[6] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[7] Y. Bazi, L. Bruzzone, and F. Melgani, "Automatic identification of the number and values of decision thresholds in the log-ratio image for change detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 3, pp. 349–353, Jul. 2006.

[8] L. I. Kuncheva and W. J. Faithfull, "PCA feature extraction for change detection in multidimensional unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 1, pp. 69–80, Jan. 2014.

[9] Q. Liu, G. Liu, C. Huang, S. Liu, and J. Zhao, "A tasseled cap transformation for Landsat 8 OLI TOA reflectance images," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 541–544, doi: 10.1109/IGARSS.2014.6946479.

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

[11] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2070–2082, Jul. 2008.

[12] W. Feng, H. Sui, J. Tu, W. Huang, and K. Sun, "A novel change detection approach based on visual saliency and random forest from multitemporal high-resolution remote-sensing images," *Int. J. Remote Sens.*, vol. 39, no. 22, pp. 7998–8021, Jul. 2018.
[13] G. Verdier and A. Ferreira, "Adaptive Mahalanobis distance and K-nearest neighbor rule for fault detection in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 1, pp. 59–68, Feb. 2011.
[14] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.
[15] X. Jiang, S. Xian, M. Wang, and P. Tang, "Dual-pathway change detection network based on the adaptive fusion module," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 8018905.
[16] Y. Wen, X. Ma, X. Zhang, and M.-O. Pun, "GCD-DDPM: A generative change detection model based on difference-feature-guided DDPM," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Mar. 2024, Art. no. 5404416.
[17] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 8006705.
[18] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067, doi: 10.1109/icip.2018.8451652.
[19] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understanding*, vol. 187, Oct. 2019, Art. no. 102783.
[20] 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.
[21] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 8007805.
[22] K. Jiang, J. Liu, W. Zhang, F. Liu, and L. Xiao, "MANet: An efficient multidimensional attention-aggregated network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 4706118.
[23] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607514.
[24] J. Ma, J. Duan, X. Tang, X. Zhang, and L. Jiao, "EATDer: Edge-assisted adaptive transformer detector for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens*, vol. 62, Dec. 2024, Art. no. 5602015.
[25] J. Xie, F. Gao, X. Zhou, and J. Dong, "Wavelet-based bi-dimensional aggregation network for SAR image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, Jul. 2024, Art. no. 4013705.
[26] Y. Wu et al., "CSTSUNet: A cross Swin transformer-based Siamese U-shape network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens*, vol. 61, Oct. 2023, Art. no. 5623715, doi: 10.1109/tgrs.2023.3326813.
[27] H. Zhang, Z. Lin, F. Gao, J. Dong, Q. Du, and H.-C. Li, "Convolution and attention mixer for synthetic aperture radar image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Sep. 2023, Art. no. 4012105.
[28] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "WNet: W-shaped hierarchical network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5615814.
[29] J. Wang, F. Gao, J. Dong, S. Zhang, and Q. Du, "Change detection from synthetic aperture radar images via graph-based knowledge supplement network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1823–1836, Feb. 2022.
[30] R. Hang, S. Xu, P. Yuan, and Q. Liu, "AANet: An ambiguity-aware network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 2024, Art. no. 5612911.
[31] Y. Liu, F. Zhang, S. Zhang, K. Zhang, J. Sun, and L. Bruzzone, "Content-guided spatial–spectral integration network for change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Jan. 2024, Art. no. 5604816.
[32] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 4401015.
[33] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Mar. 2022, doi: 10.1007/s41095-022-0274-8.
[34] C.-P. Chen, J.-W. Hsieh, P.-Y. Chen, Y.-K. Hsieh, and B.-S. Wang, "SARAS-Net: Scale and relation aware Siamese network for change detection," in *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, pp. 14187–14195, doi: 10.1609/aaai.v37i12.26660.
[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–18.
[36] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, May 2022.
[37] 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, Apr. 2022, Art. no. 4410213.
[38] C. Han, C. Wu, H. Guo, M. Hu, J. Li, and H. Chen, "Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8395–8407, Aug. 2023.
[39] Z. Zhang, L. Bao, S. Xiang, G. Xie, and R. Gao, "B2CNet: A progressive change boundary-to-center refinement network for multitemporal remote sensing images change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 37, pp. 11322–11338, Jun. 2024.
[40] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[41] 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, May 2020, Art. no. 1662.
[42] 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, Jun. 2021, Art. no. 5604816.
[43] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.
[44] M. Liu, S. Lin, Y. Zhong, Q. Shi, and J. Liu, "A memory-guided network and a novel dataset for cropland semantic change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024, Art. no. 4410013.

Chuang Liu is currently working toward the M.Eng. degree in computer technology with the School of Computer Science, Hubei University of Technology, Wuhan, China.

His research interests include remote sensing image processing, multimodal image fusion, low-level vision, machine learning, and deep learning.

Liyang Bao received the B.Sc. degree in computer science and technology from Shangqiu University, Shangqiu, China, in 2022. He is currently working toward the M.Eng. degree in computer technology with the School of Computer Science, Hubei University of Technology, Wuhan, China.

His research interests include intelligent remote sensing image processing, change detection, and deep learning.

Zhiqi Zhang received the B.Sc. degree in geographic information system from Huazhong Agricultural University, Wuhan, in 2006, the B.Eng. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, in 2006, and the M.Eng. degree in computer technology and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2015 and 2018, respectively.

He is currently an Associate Professor with the School of Computer Science, Hubei University of Technology, Wuhan. His research interests include system architecture, algorithm optimization, AI, and high-performance processing of remote sensing.