# Transformer With Feature Interaction and Fusion for Remote Sensing Image Change Detection

Dongen Guo, Tao Zou, Ying Xia, *Member, IEEE*, and Jiangfan Feng

***Abstract*—With the rapid development of deep learning (DL), change detection (CD) in remote sensing (RS) image has achieved remarkable success. Nevertheless, as the image resolution improves, the visual features extracted by current methods have limited expression ability, and the networks generally suffer from spatial degradation, both of which lead to incomplete boundary detection and the undetection problem of small changed areas. At the same time, the registration errors in image pairs make remote sensing image change detection (RSCD) more challenging. To alleviate the aforementioned issues, this article proposes a transformer with feature interaction and fusion network (TFIFNet) for CD. To be specific, the proposed network utilizes the advantages of transformer in long-range dependence modeling first, which can learn feature representations with spatial-temporal information from a global perspective. Then, to alleviate the irrelevant changes caused by image registration errors, the bitemporal feature interaction module (BFIM) is proposed, which utilizes an attention mechanism to learn the bitemporal background distribution. Subsequently, an intertemporal joint-attention (JointAtt) is introduced to learn the consistency of bitemporal features for further refinement. Finally, to address the issue caused by spatial degradation during the process of network training, a triple feature fusion module (TFFM) is proposed. This module can learn spatial information from adjacent layer's features as additional spatial information. Extensive experimental studies show that the proposed network achieves the most advanced results on two CD benchmark datasets.**

***Index Terms*—Attention mechanism, change detection (CD), deep learning (DL), remote sensing (RS), transformer.**

Received 26 April 2024; revised 29 July 2024; accepted 17 August 2024. Date of publication 26 August 2024; date of current version 6 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 41971365 and Grant 41571401, in part by the Science and Technology Research Project of Henan Province under Grant 212102210492, in part by the Key Research Projects of Henan Higher Education Institutions under Grant 23A520053, in part by the Doctoral Research Start-up Fund Project at Nanyang Institute of Technology under Grant NGBJ-2023-32, and in part by the Interdisciplinary Sciences Project of Nanyang Institute of Technology under Grant NGJC-2022-01. *(Corresponding authors: Dongen Guo; Tao Zou.)*

Dongen Guo is with the Key Lab of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China, and also with the School of Computer and Software, Nanyang Institute of Technology, Nanyang 473004, China (e-mail: gden_2008@126.com).

Tao Zou and Ying Xia are with the Key Lab of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: zout_99@foxmail.com; xiaying@cqupt.edu.cn).

Jiangfan Feng is with the Chongqing Engineering Research Center for Spatial Big Data Intelligent Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: fengjf@cqupt.edu.cn).

Data is available on-line at https://github.com/liumency/CropLand-CD and https://github.com/liumency/SYSU-CD.

Digital Object Identifier 10.1109/JSTARS.2024.3449923

## I. Introduction

REMOTE sensing image change detection (RSCD) is the process of comparing and recognizing the differences of the same location at different times, which is the key to understanding land surface change and land surface activity detection [1]. As one of the key research topics in the field of computer vision (CV), it has been widely used in urban planning [2], disaster management [3], land use [4], and environmental monitoring [5]. With the rapid development of sensors, the massive high-resolution RS data urgently require highly automated CD methods to reduce the cost of manual interpretation. As a result, automatic CD has gradually emerged in the field of RSCD and become an important research topic due to its high efficiency and accuracy, which provides strong support for land surface CD.

In the early phases of RSCD, traditional image processing technologies, such as principal component analysis (PCA) [6], change vector analysis (CVA) [7], were widely introduced. However, the methods based on traditional technologies have a significant disadvantage, that is, their performance depends heavily on the selection of thresholds. Then, machine learning-based methods have garnered considerable attention, methods such as support vector machine (SVM) [8] and random forest [9] have begun to be introduced. However, as image resolution gradually improved, machine learning-based methods generally exhibit poor generalization in complex scenarios.

Deep learning (DL)-based models, which possess nonlinear mapping capabilities that enable them to capture intricate image details and complex texture features, have already made significant achievements in fields such as image analysis [10], [11], [12], natural language processing (NLP) [13], [14], [15], image scene classification [16], [17], [18]. Convolutional neural network (CNN) has attracted particular attention due to its functionality of automatically capturing complex and nonlinear features, and has made outstanding achievements in object detection [19], [20], image classification [21], [22], semantic segmentation [23], [24], face recognition [25], [26], and other fields. Following this trend, CD in RS image has developed rapidly. Numerous outstanding CD networks based on CNN have been proposed [27], [28], [29]. For example, Zhan et al. [27] introduced the Siamese convolutional network into CD tasks to efficiently process bitemporal images simultaneously for the first time. Daudt et al. [28] proposed Siamese architecture for end-to-end training of RSCD. Daudt et al. [29] later proposed three well-known CD networks that based on fully convolutional neural networks (FCNs).

Although CNN-based models have achieved promising results, these models are inherently constrained by the size of receptive fields [30], which extremely hinders their abilities to effectively extract global spatial-temporal information that is crucial for recognizing changes in RSCD [31]. To address this issue, several approaches have been explored and adopted, including stacking convolutional layers [32], [33] to deepen the architecture of networks, using dilated convolution [34] and using attention mechanisms [35], [36]. Although these methods have indeed broadened the receptive field of CD networks to some degree, simply using attention mechanism to fuse or weight features in the spatial or channel dimensions fails to capture long-range dependencies effectively.

Transformer [37] was proposed in 2017 and originally applied to NLP tasks. Compared with traditional CNN, transformer overcomes the limitation of receptive fields, providing a novel solution for CV tasks. Therefore, scholars began to introduce transformer into CV tasks, and remarkable performance has been achieved in tasks such as object detection [38], [39], semantic segmentation [40], [41], and image classification [42], [43]. With the development of DL, some excellent transformer architectures have emerged, including Vision Transformer [44], SegFormer [40], and Swin Transformer [45]. These transformers can further provide more powerful context modeling capabilities than CNN. However, in the specific field of RSCD, the transformer is often used as a part of feature processing [1], [31], [46], [47], and such methods have not yet fully leverage the advantages of transformer in global feature learning and spatial-temporal modeling.

Although current models have made notable advancements in the field of RSCD, there are still some issues to be solved. First, the time span of RS image sampling is generally large, and the positions during sensor sampling are uncertain, which often results in spatial, illumination, and seasonal differences between images sampled at the same location but at different times. This objective difference, that is, registration error, makes the model prone to false detection easily. Second, in the forward processing of the model, the features become more and more abstract. Although the semantic information of the features is richer, a lot of detailed information is lost, which often makes the model confused when locating the boundary, resulting in missed detection, especially for small changed regions. Finally, the feature fusion method adopted by most of the current methods ignores the semantic gap between low-level features and high-level features, which makes it easy to confuse the network. These issues mentioned above have seriously hindered the further development of RSCD.

In order to alleviate the problems mentioned above, we propose a transformer with feature interaction and fusion network (TFIFNet) for CD. Following the current trends in the CD field, TFIFNet adopts the Siamese architecture and uses pretrained Swin transformer as the backbone. First, the multilevel features of bitemporal images are extracted. Then, aiming at distinguishing relevant changes from irrelevant changes caused by registration errors to obtain more discriminative features, the bitemporal feature interaction module (BFIM) is introduced to realize feature interaction between bitemporal images. Finally, to improve boundary integrity and alleviate the problem of missed detection in small changed area, a triple feature fusion module (TFFM) integrated with intertemporal joint-attention (JointAtt) is proposed, which can further suppress irrelevant changes and mitigate the impact of semantic gap while reducing the problem of spatial information loss. The key contributions of this work are presented below.

1) An innovative network named TFIFNet is proposed for RSCD, which includes two main components, namely, BFIM and TFFM. It can significantly reduce the false detection caused by registration errors, effectively alleviate the problem of information loss in the network's feed-forward process, which leads the detection accuracy and stability of the network.
2) The BFIM employees spatial attention to obtain the spatial distribution, and then accurately captures the spatial context by learning the background distribution in the bitemporal images to obtain more discriminative features, thereby improving the robustness of the network against registration errors.
3) The TFFM is designed to address the issue of information loss and avoid the impact of semantic gap, as it integrates the bitemporal features of adjacent layer serving as complementary spatial information rather than multilevel features.
4) Experimental results on two widely used RSCD benchmarks have confirmed the effectiveness of the proposed network TFIFNet.

The rest of this article is organized as follows. Works related to this research are described in Section II. The details of the proposed TFIFNet are described in Section III. The comprehensive experimental evaluation is carried out in Section IV. Finally, Section V concludes the article.

## II. Related Work

### A. Feature Interaction-Based Methods

Feature interaction is mainly used in machine learning and data analysis, which can capture the complex relationships in data. Unlike other dense prediction tasks, the interaction between bitemporal features is a worthwhile factor to consider in CD tasks. Based on this consideration, Fang et al. [48] proposed a general architecture for CD named MetaChanger, which incorporates a sequence of optional interaction layers in its feature extractor. They also defined the concept of feature interaction in CD as the correlation or communication among homo/heterogeneous features during the extraction phase prior to fusion. Subsequently, Liu et al. [31] introduced the parameter and computation-free operations in MetaChanger to make the distributions between dual branches more similar by exchanging bitemporal features, thus helping the model to bridge the domain gaps. There are also many other models based on feature interaction in the field of RSCD as well, for instance, Song et al. [1] proposed the dual-feature mixed attention module for the first time, a novel approach that leverages distinct coarse-grained features for information interaction to strike a balance between the neglect caused by excessive deep sampling and the

blurred edges resulting from insufficient shallow sampling. Lu et al. [49] used relational cross-attention to acquire bitemporal relationship-aware features, thereby acquiring pixel embeddings with rich global information.

Based on the idea of feature interaction, the mentioned models above mine the internal relationship between features, which effectively improves the performance of CD models. However, most of these models rely heavily on hybrid attention, which will increase the computational complexity significantly. The feature interaction method proposed by Fang et al. [48] that exchanged some bitemporal features in the channel or spatial dimension may lose some important information, resulting in incomplete features and subsequently affecting the performance of the model. To mitigate such issue, we only exchange the background distribution based on spatial attention map to maintain the integrity of features, rather than exchanging the bitemporal features.

### B. Feature Fusion-Based Methods

Feature fusion is crucial to the success of DL-based models, as it can produce more discriminative representations by combining features from multiple layers or different source data [50]. In current DL-based methods, most of them regard the fusion of multilevel features as an important way to improve the performance of models or networks, since multilevel integration enables the aggregation of both spatial details and semantic information. The field of RSCD also continues this trend. For example, Bandara et al. [51] achieved the integration of different features across various scales through cascades and multilayer perceptrons (MLP). Liu et al. [31] used elementwise summation to achieve dual-branch multilevel feature fusion. Fang et al. [33] used dense skip connections to achieve multilevel feature fusion in the decoder stage. Zhang et al. [32] used channel and spatial attention to integrate difference features and bitemporal features for change map reconstruction. Xu et al. [52] utilized MLP and other techniques to fuse features extracted from CNN and Transformer branches in the decoder stage.

The methods mentioned above have achieved good results through the feature fusion, but the semantic discrepancy between low-level and high-level features is ignored, which will introduce discrepancies and confusion into the networks when these features are fused directly [53]. It is also worth noting that most models directly fuse multiscale bitemporal features [49], [54], [55]. However, due to registration errors, when fusing bitemporal features, the change features within one temporal may become contaminated by background features at the same spatial location in another time series, resulting in poorer network performance [56]. Overcoming these potential issues is crucial for enhancing the performance of CD models. Inspired by Ye et al. [53], we incorporate features from adjacent layer into the fusion process of each layer instead of multiscale features.

### C. Transformer-Based Methods

Transformer was proposed in 2017, and it has garnered remarkable results in NLP. With the successful application of transformer in image classification [43], [57], semantic segmentation [40], [41], object detection [38], [39], and other CV fields, transformer has gradually become one of the mainstream methods for CV tasks. The amazing performance of transformer in both NLP and CV tasks has attracted the interest of the RS community, thus, researchers begin to introduce transformer into RSCD. For example, Song et al. [1] optimized the extracted semantic tokens using transformer and fed them back into the original features to reconstruct the pixel space. Jiang et al. [46] leveraged transformer to mine the relationships between tokens representing invariant backgrounds, providing clues for learning the consistency of bitemporal images. Li et al. [58] made the first attempt to combine transformer and UNet, and used transformer to learn the global context to further enhance the representation ability of features extracted by CNN. Liu et al. [31] also employed transformer later to effectively model the contextual information of bitemporal features extracted by CNN. Lu et al. [49] utilized transformer to explore the long-range contextual information and correspondence of bitemporal features. The methods mentioned above have achieved good results by introducing transformer, but they still exhibit the following shortcoming. Specifically, they fail to fully leverage the transformer's capabilities in multilevel feature learning and instead treat it merely as a tool to enrich the contextual information of bitemporal features.

Different from most of the existing transformer-based CD networks, the proposed network utilizes transformer for global feature extraction rather than as a feature processing tool. TFIFNet also mines the background distribution of bitemporal images through feature interaction, but retains the integrity of the features. It is also worth noting that this network adopts a different fusion strategy, which uses adjacent layer features to fuse with other features of the current layer to alleviate the impact of the semantic gap and alleviate the spatial degradation problem.

## III. Methodology

### A. Overview

The architecture of the proposed TFIFNet is illustrated in Fig. 1, which provides a new and effective solution for the RSCD tasks by interactively learning the background distribution through bitemporal features. TFIFNet mainly contains three parts: the BFIM, the JointAtt, and the TFFM, where the JointAtt is integrated in the TFFM. For any input image pairs, multilevel bitemporal visual features are first extracted from a global perspective in parallel through two shared Swin transformer branches. Subsequently, the BFIM is employed to learn the background distribution of bitemporal features within the deep encoder, which can guide the two branches to learn the consistency of the background in bitemporal images, thereby suppressing the impact of irrelevant changes caused by registration errors. Then, the global feature distribution of each input is guided by JointAtt to further suppress irrelevant changes so as to better capture the change features, and the refined bitemporal features are sent to the TFFM. In TFFM, the adjacent layer's features, the refined bitemporal features, and the difference features are fused through concatenation along the channel dimension. This fusion process not only compensates for the issue

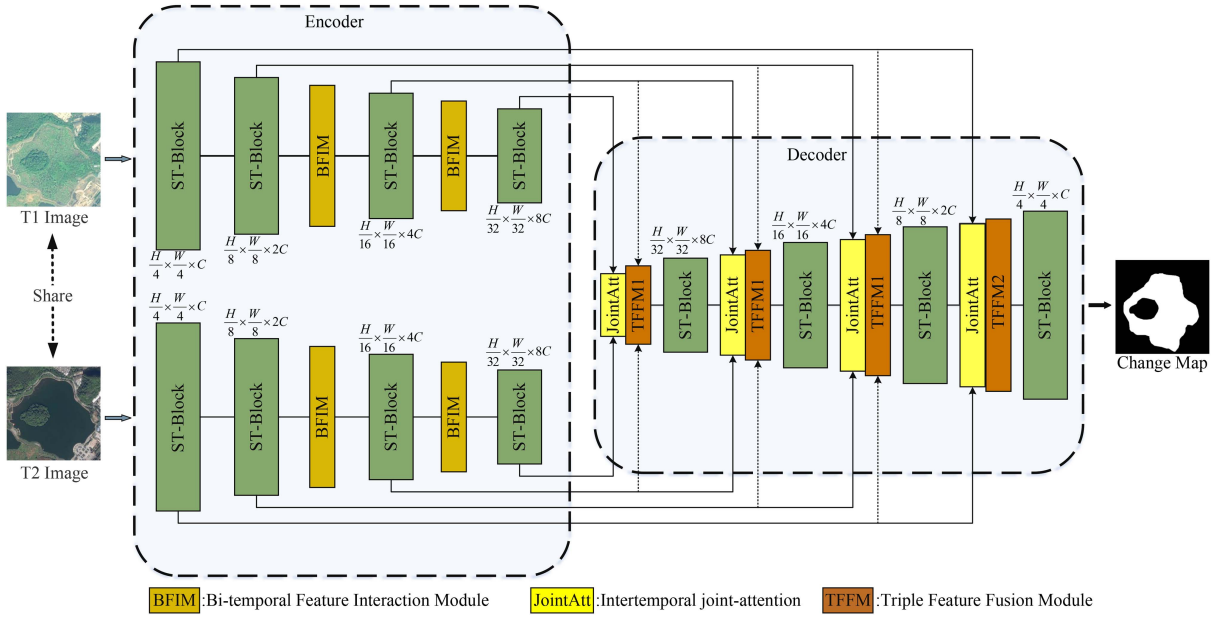


Fig. 1. Architecture of the proposed TFIFNet. The pretrained Swin Transformers are first used to extract the bitemporal features, and the BFIM is used to suppress irrelevant changes. Then, the intertemporal JointAtt module is used to further refine the bitemporal features, and finally, the TFFM is used to fusion features, including bitemporal features, difference features, and adjacent features.

of network space degradation but also corrects the difference features. TFIFNet emphasizes highlighting the changed regions through the distribution of background information to achieve precise CD results, rather than relying directly on bitemporal features or difference features.

### B. Bitemporal Feature Interaction Module

By observing the relevant datasets and papers, we find that the proportion of background is often higher than that of changed regions in the bitemporal images. Unfortunately, most of the current networks mainly focus on refining the change features or refining the difference features to obtain better detection results, ignoring the mining of rich background information in bitemporal images. Learning the background information is beneficial for CD tasks, because the inconsistent imaging conditions of RS sensors will lead to the differences in angle and illumination intensity during sampling, and such differences may lead to occlusion of true changes, which is also a great challenge for current CD tasks. Jiang et al. [46] have confirmed that mining the common background information carefully can serve as a crucial cue for learning consistent representations between the two images. Meanwhile, the essence of CD is to locate changes by comparing semantic differences. But the semantic differences in unchanged areas are often small, the semantic differences in changed areas are often large, and the location of the changed areas is also uncertain.

Based on the aforementioned two facts, we have the intuition that the distribution of background information in the bitemporal images should be similar, and by learning the consistency of the background in the bitemporal images, the changed regions can be easily represented, and the irrelevant changes can be ignored as well. Influenced by the feature interaction concept of Fang et al. [48], we propose BFIM. But unlike Fang et al. [48], we preserve the integrity of bitemporal features just by exchanging the background distributions of bitemporal images. To avoid the impact of irrelevant information in the background, we also leverage residual learning to ensure that the network is robust against such variations. The detailed architecture of BFIM is depicted in Fig. 2.

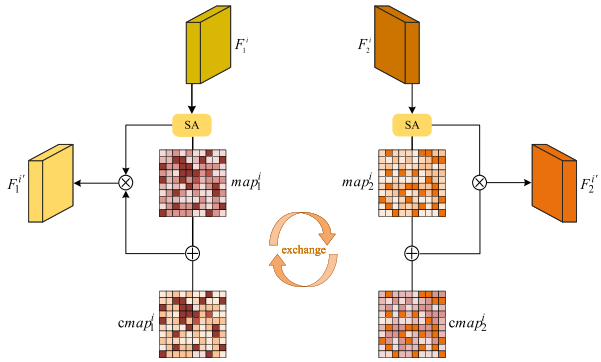


Fig. 2. Illustration of the BFIM.

Specifically, the input bitemporal images $T_1 \in R^{H\times W\times C}$ and $T_2 \in R^{H\times W\times C}$ are fed into the weight shared encoder to obtain bitemporal features $F_1^i$ and $F_2^i$ first, where i denotes the index of each layer in encoder. To effectively learn the background distribution in bitemporal features to better distinguish the background from the foreground, and alleviate the impact of irrelevant changes caused by registration errors, the BFIM is inserted before the third and fourth layers, respectively. We designed it this way because the features extracted in the very shallow layer are relatively simple, which cannot accurately reflect the complex relationship between the background and the foreground.

**Algorithm 1:** BFIM, PyTorchlike Code.

```
import torch
import torch.nn as nn

class BFIM(nn.Module):
    def __init__(self):
        super().__init__()
        self.sam = SpatialAttention(kernel_size=3)

    def forward(self, x1, x2):
        map1 = self.sam(x1)
        map2 = self.sam(x2)
        avg_weight1 = torch.mean(map1)
        avg_weight2 = torch.mean(map2)
        low_weight1 = map1 < avg_weight1
        low_weight2 = map2 < avg_weight2
        exchange_map1 = map1.clone
        exchange_map2 = map2.clone
        exchange_map1[low1] = map2[low1]
        exchange_map2[low2] = map1[low2]
        out1 = x1 * (exchange_map1 + map1)
        out2 = x2 * (exchange_map2 + map2)

        return out1, out2
```

The features in the very deep layer are highly abstract and the degradation of spatial information is serious. To guarantee the efficacy of the extracted background information and make sure it still retains part of the spatial information, we choose to insert the modules in the middle layer, and experiment in Section IV, Table V also proves that this is beneficial. For bitemporal feature pair $(\mathrm{F}_1^{\mathrm{i}}, \mathrm{F}_2^{\mathrm{i}})$, the spatial attention mechanism is first used to obtain the attention maps $\mathrm{map}_1^{\mathrm{i}}$ and $\mathrm{map}_2^{\mathrm{i}}$. Then, $\mathrm{low\_weight}_{\mathrm{j}}^{\mathrm{i}}$ of the region with weights below the average weights $\mathrm{avg\_weight}_{\mathrm{j}}^{\mathrm{i}}$ are calculated in $\mathrm{map}_1^{\mathrm{i}}$ and $\mathrm{map}_2^{\mathrm{i}}$, respectively, to obtain the background distribution. The procedure can be written as

$$\mathrm{low\_weight}_{\mathrm{j}}^{\mathrm{i}} = \mathrm{map}_{\mathrm{j}}^{\mathrm{i}} < \mathrm{avg\_weight}_{\mathrm{j}}^{\mathrm{i}},\ \mathrm{j} = 1, 2. \tag{1}$$

After obtaining the regions $\mathrm{low\_weight}_{\mathrm{j}}^{\mathrm{i}}$, we obtain the attention map $\mathrm{cmap}_{\mathrm{j}}^{\mathrm{i}}$ that includes another temporal's background distribution by exchanging the weights of the corresponding index representation. In order to avoid the background information being polluted by irrelevant factors, $\mathrm{cmap}_{\mathrm{j}}^{\mathrm{i}}$ and $\mathrm{map}_{\mathrm{j}}^{\mathrm{i}}$ are added with residuals, followed by elementwise multiplication with the raw features $\mathrm{F}_{\mathrm{j}}^{\mathrm{i}}$. Finally, the refined bitemporal feature pairs $(\mathrm{F}_1^{\mathrm{i}'}, \mathrm{F}_2^{\mathrm{i}'})$ are obtained. The pseudocode of BFIM is in Algorithm 1.

### C. Intertemporal JointAtt

As the essence of CD tasks lies in comparing bitemporal information, it is inherently driven by cross-attention [49]. Based on this concept, we adopt the intertemporal JointAtt module (Feng et al. [59]), which integrates self-attention and cross-attention into one module to guide the overall feature distribution for each input, so as to further suppress the impact of irrelevant changes

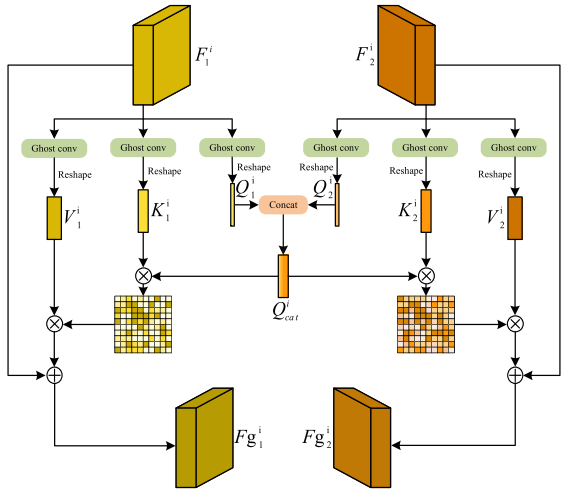


Fig. 3. Illustration of the JointAtt module.

in the bitemporal features and better capture the changed regions. The JoinAtt is shown in Fig. 3. Specifically, in order to combine the information of bitemporal features and accurately capture the semantic differences between different temporal to suppress irrelevant changes, JointAtt integrates the partial representation of another temporal while retaining the feature information of the current branch.

It is worth noting that due to the relatively high computational complexity of self-attention and cross-attention, the traditional convolution mapping method used in the original JointAtt will further increase the complexity of the network. To address this challenge, we introduce model compression techniques to replace traditional convolutions in JointAtt with ghost convolutions [60]. It is also worth noting that Feng et al. [59] integrated the JointAtt in the encoder stage, while we integrated JointAtt in the decoder stage, because our ultimate goal is to correct the difference feature through the refined bitemporal features, since the difference features lack the temporal information, irrelevant changes that are amplified during the process of obtaining the difference features may affect the localization of changed regions.

More precisely, the bitemporal features $\mathrm{F}_1^{\mathrm{i}}$ and $\mathrm{F}_2^{\mathrm{i}}$ are first mapped via a 1x1 Ghost convolution layer and subsequently reshaped into embeddings that encompass query $\mathrm{Q}_{\mathrm{j}}^{\mathrm{i}}$, key $\mathrm{K}_{\mathrm{j}}^{\mathrm{i}}$, and value $\mathrm{V}_{\mathrm{j}}^{\mathrm{i}}$. Then, $\mathrm{Q}_1^{\mathrm{i}}$ and $\mathrm{Q}_2^{\mathrm{i}}$ are concatenated along the channel dimension into $\mathrm{Q}_{\mathrm{cat}}^{\mathrm{i}}$, which contains two temporal's information. Next, the similarity between $\mathrm{Q}_{\mathrm{cat}}^{\mathrm{i}}$ and $\mathrm{K}_1^{\mathrm{i}}$ or $\mathrm{K}_2^{\mathrm{i}}$ can be obtained through a sequential process involving multiplication and Softmax operations. Next, the attention map is obtained by weighted summation with $\mathrm{V}_{\mathrm{j}}^{\mathrm{i}}$ using the similarity. Finally, the original input $\mathrm{F}_{\mathrm{j}}^{\mathrm{i}}$ is integrated via a skip connection to obtain further rectified features $\mathrm{Fg}_{\mathrm{j}}^{\mathrm{i}}$. Mathematically, the overall process can be formulated as follows:

$$\begin{aligned}\mathrm{Fg}_{\mathrm{j}}^{\mathrm{i}} &= \mathrm{JointAtt}\left(\{\mathrm{Q}_1^{\mathrm{i}}, \mathrm{Q}_2^{\mathrm{i}}\}, \{\mathrm{K}_{\mathrm{j}}^{\mathrm{i}}, \mathrm{V}_{\mathrm{j}}^{\mathrm{i}}\}\right) \\ &= \mathrm{Softmax}(\mathrm{Concat}\left(\mathrm{Q}_1^{\mathrm{i}}, \mathrm{Q}_2^{\mathrm{i}}\right) \cdot \mathrm{K}_{\mathrm{j}}^{\mathrm{i}}) \cdot \mathrm{V}_{\mathrm{j}}^{\mathrm{i}} + \mathrm{F}_{\mathrm{j}}^{\mathrm{i}} \\ &= \mathrm{Softmax}\left(\mathrm{Q}_{\mathrm{cat}}^{\mathrm{i}}, \mathrm{K}_{\mathrm{j}}^{\mathrm{i}}\right) \cdot \mathrm{V}_{\mathrm{j}}^{\mathrm{i}} + \mathrm{F}_{\mathrm{j}}^{\mathrm{i}}\end{aligned} \tag{2}$$

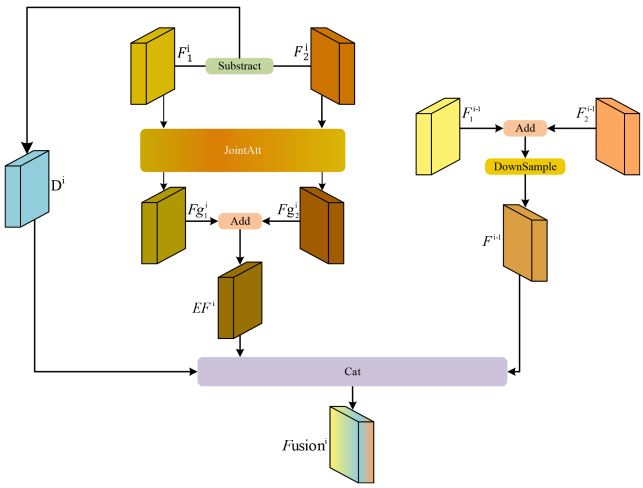


Fig. 4. Illustration of the TFFM.

where $i = 1, 2, 3, 4$, notes the encoder stage, and $j = 1, 2$, represents the serial number of different temporal. By preserving its own features while introducing representations from another input, JointAtt effectively utilizes spatial-temporal contextual features, further enhancing the global attention to focus more on the true regions of changes.

### D. Triple Feature Fusion Module

To achieve optimal detection results, the current CD tasks mainly focus on extracting fine-grained bitemporal features to obtain difference features that can accurately locate the changed regions. However, during the extraction of bitemporal features, these methods often prioritize the extraction of deep semantic features while neglecting the high-resolution and fine-grained shallow information, leading to uncertainties in the edge pixels of the changed regions and the absence of identification for small targets in the extracted features. To address these issues, a straightforward approach is to directly fuse features from different levels. Nevertheless, due to the semantic discrepancy between high-level and low-level features, direct fusion can introduce discrepancies and confusion in the networks. Furthermore, as the difference features do not contain bitemporal information, they are susceptible to irrelevant variations.

To mitigate these challenges, we propose TFFM, as depicted in Fig. 4. This module compensates for the absence of bitemporal information in difference features by leveraging bitemporal features and alleviates the impact of the semantic gap through the fusion of adjacent layer's features. This fusion approach fully utilizes bitemporal features, difference features, and adjacent features, providing a novel fusion strategy for enhancing the performance of CD networks.

Specifically, the bitemporal feature pair $(\mathrm{F}_1^{\mathrm{i}}, \mathrm{F}_2^{\mathrm{i}})$ is first fed into the JointAtt module to acquire the enhanced feature pair $(\mathrm{Fg}_1^{\mathrm{i}}, \mathrm{Fg}_2^{\mathrm{i}})$, followed by the elementwise addition to get $\mathrm{EF}^{\mathrm{i}}$. Next, the difference feature $\mathrm{D}^{\mathrm{i}}$ is obtained by direct subtraction of raw bitemporal feature pairs. The reason why we perform direct subtraction operations on the original features rather than on the refined bitemporal features is that the refined features tend to pay more attention to the areas of changes, rather than global information. Then, we incorporate the features from the previous layer, that is the adjacent layer's feature pair $(\mathrm{F}_1^{\mathrm{i}-1}, \mathrm{F}_2^{\mathrm{i}-1})$, as supplementary spatial information. Subsequently, an elementwise addition is performed as well to get feature $\mathrm{F}^{\mathrm{i}-1}$. Then, downsampling is applied to $\mathrm{F}^{\mathrm{i}-1}$. After getting these three features, the feature fusion is achieved by concatenating along the channel dimension. The fusion process can be formulated as

$$\mathrm{Fuse}^{\mathrm{i}} = \mathrm{Concate}\left(\mathrm{EF}^{\mathrm{i}}, \mathrm{D}^{\mathrm{i}}, \mathrm{DownSample}\left(\mathrm{F}^{\mathrm{i}-1}\right)\right). \quad (3)$$

It is noteworthy to mention that, as the fusion method of adjacent layer features is adopted, only the bitemporal features and the corresponding difference features are fused in the final stage of the decoder.

### E. Loss Function

To enhance the performance of our proposed TFIFNet, we adopt the loss function in [61], which is the sum of weighted binary cross-entropy loss, structural similarity loss, and soft intersection over union loss. The overall loss is expressed in (4). The $\mathcal{L}$wbce can effectively addresses the imbalance issue associated with change pixels, the $\mathcal{L}$ssim emphasizes the local structure of change boundaries, and the $\mathcal{L}$siou pays more attention to the changed regions.

$$\mathcal{L} = \sum_{s=1}^{S} \mathcal{L}\mathrm{wbce}^s + \mathcal{L}\mathrm{ssim}^s + \mathcal{L}\mathrm{siou}^s \quad (4)$$

where S denotes the total layer of the encoder, in our work, the total number of layers is 4, i.e., $\mathrm{S} = 4$. By paying more attention to the changed regions, TFIFNet can achieve more precise CD results when optimizing the aforementioned loss.

## IV. Experiments

### A. Datasets

We use two CD datasets to thoroughly validate the effectiveness of the proposed network TFIFNet, they are CLCD [62], SYSU-CD [63].

CropLand Change Detection [62] contains 600 pairs of farmland change sample images, 320 pairs are used for training, 120 pairs are used for validation, and 120 pairs are used for testing. The bitemporal images of CLCD were captured by the Gaofen-2 satellite in Guangdong Province, China, in 2017 and 2019, respectively, possessing a spatial resolution ranging from 0.5 to 2 m. Each set of samples is composed of two images of $512 \times 512$ pixels and the corresponding binary change labels. The main types of changes noted in the CLCD include buildings, roads, lakes, and bare soil land. We cut the images along the center with nonoverlapping to the size of $224 \times 224$, and obtain 1440/480/480 samples for train/val/test for these images, respectively.

The SYSU-CD [63] dataset contains 20 000 pairs of aerial images captured in Hong Kong from 2007 to 2014 with a resolution of 0.5 m and a size of $256 \times 256$. Among these images, 12 000 pairs for training, 4000 pairs for validation, and another 4000 pairs for testing. The primary types of changes

annotated in this dataset include: 1) newly constructed urban buildings; 2) suburban expansion; 3) preparatory work before construction; 4) vegetation changes; 5) road expansions; and 6) maritime constructions. We performed nonoverlapping cutting along the center of each image pair, resulting in image pairs with a size of 224 × 224, 20 000/4000/4000 for train/val/test, respectively.

### B. Implementation Details

The proposed network TFIFNet was implemented within the PyTorch framework and trained in an environment powered by the Ubuntu operating system, with acceleration provided by GeForce RTX 3090. We utilized the mini-batch SGD algorithm to train the network, with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.0005. Due to the significant disparity in the number of samples across the two datasets, we established distinct batch sizes for the two datasets, respectively, which is 6 for the CLCD dataset and 10 for the SYSU-CD dataset. For the backbone of the Siamese network, we utilized the Swin transformer that had been pretrained on the ImageNet-22k classification task [64]. For the remaining layers, we initialized them randomly and assigned a learning rate 10 times higher than the original one. The network was trained for 80 epochs. The learning rate decreases to 1/10 of the initial learning rate every 20 epochs.

### C. Evaluation Criteria

To assess the performance of our proposed network, we employ the F1 score and the intersection over union ratio (IoU) as the primary evaluation metrics. And the precision, recall, and overall precision (OA) are used as auxiliary metrics. Among the evaluation metrics, F1-score, IoU, and OA serve as comprehensive indicators, and larger values indicate better prediction. Each metric is defined as follows:

$$\begin{aligned} \text{Precision} &= \text{TP}/(\text{TP}+\text{FP}) \\ \text{Recall} &= \text{TP}/(\text{TP}+\text{FN}) \\ \text{F1} &= 2\cdot\text{Precision}\cdot\text{Recall}/(\text{Precision}+\text{Recall}) \\ \text{IoU} &= \text{TP}/(\text{TP}+\text{FN}+\text{FP}) \\ \text{OA} &= (\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{FP}+\text{TN}) \end{aligned} \tag{5}$$

where TP, FP, FN, and TN are the numbers of true positives, false positives, false negatives, and true negatives, respectively.

### D. Comparison to State of the Art

We have compared TFIFNet with several excellent methods, they are FC-Siam-conc [29], FC-Siam-diff [29], DMINet [59], SEIFNet [65], VcT [46], FTN [61], ChangeFormer [51], BIT [47], and ICIFNet [30]. A brief introduction to these models is listed below.

1) *FC-Siam-conc [29]:* It is a Siamese network, and the encoder consists of two FCN for extracting features in parallel. The bitemporal information is fused through feature cascade within the decoder.

TABLE I
COMPARISON RESULTS ON CLCD DATASET

| Method | Pre | Rec | F1 | IoU | OA |
|---|---|---|---|---|---|
| FC-Siam-conc [29] | 82.49 | 79.15 | 80.71 | 70.96 | 94.96 |
| FC-Siam-diff [29] | 80.80 | 79.05 | 79.89 | 70.01 | 94.62 |
| DMINet [59] | 81.93 | 78.39 | 80.03 | 70.20 | 94.80 |
| SEIFNet [65] | 77.78 | 78.82 | 78.29 | 68.17 | 93.91 |
| VcT [46] | 80.58 | 73.61 | 76.55 | 66.55 | 94.27 |
| FTN [61] | 89.14 | 82.18 | 85.25 | 76.45 | 96.16 |
| ChangeFormer [51] | 80.00 | 82.05 | 80.98 | 71.17 | 94.58 |
| BIT [47] | 81.45 | 80.39 | 80.91 | 71.15 | 94.83 |
| ICIFNet [30] | 85.17 | 77.59 | 80.83 | 71.16 | 95.28 |
| TFIFNet | **90.47** | **83.95** | **86.86** | **78.55** | **96.55** |

All the scores are described in percentage (%).
The bold face values indicate the best performing values on a single metric.

2) *FC-Siam-diff [29]:* It is a U-shaped structure, just like FC-Siam-conc. The key difference lies in the fact that FC-Siam-diff integrates the bitemporal characteristics through absolute difference approach.
3) *DMINet [59]:* It is a Siamese network, which uses self-attention and cross-attention to suppress irrelevant changes in CNN extracted features, and captures difference features from two branches of pixel-level subtraction and channel-level connection.
4) *SEIFNet [65]:* It is a CNN-based Siamese network that aims to alleviate the problems of pseudo changes and scale variations through spatiotemporal enhancement and interlevel fusion.
5) *VcT [46]:* It leverages both intra-image and inter-image cues by effectively capturing the dependencies among reliable tokens present in the dual images, and it is a transformer-based Siamese network.
6) *FTN [61]:* It is a transformer-based Siamese framework, which utilizes a variant of transformer, Swin transformer, to extract features from a global perspective. Meanwhile, FTN employs both feature summation and difference to enhance the feature.
7) *ChangeFormer [51]:* It is also a transformer-based Siamese network structure that integrates an MLP decoder with a hierarchically structured transformer encoder, aiming to provide accurate change information.
8) *BIT [47]:* It is a CNN-transformer-based Siamese network, which leverages the transformer to enrich the contextual information of ConvNet features through semantic tokens, followed by a feature differencing process to derive the change map.
9) *ICIFNet [30]:* It is also a CNN-transformer-based network, the two branches are composed of CNN and transformer, respectively, in order to comprehensively aggregate both local and global features. It utilizes intrascale cross interaction and interscale feature fusion.

We have implemented the aforementioned CD networks using their publicly available codes with default hyperparameters.

### E. Results and Analysis on CLCD

1) *Quantitative Analysis on CLCD:* As evident from Table I, our network TFIFNet significantly surpasses other SOTA networks on the CLCD dataset in terms of precision,

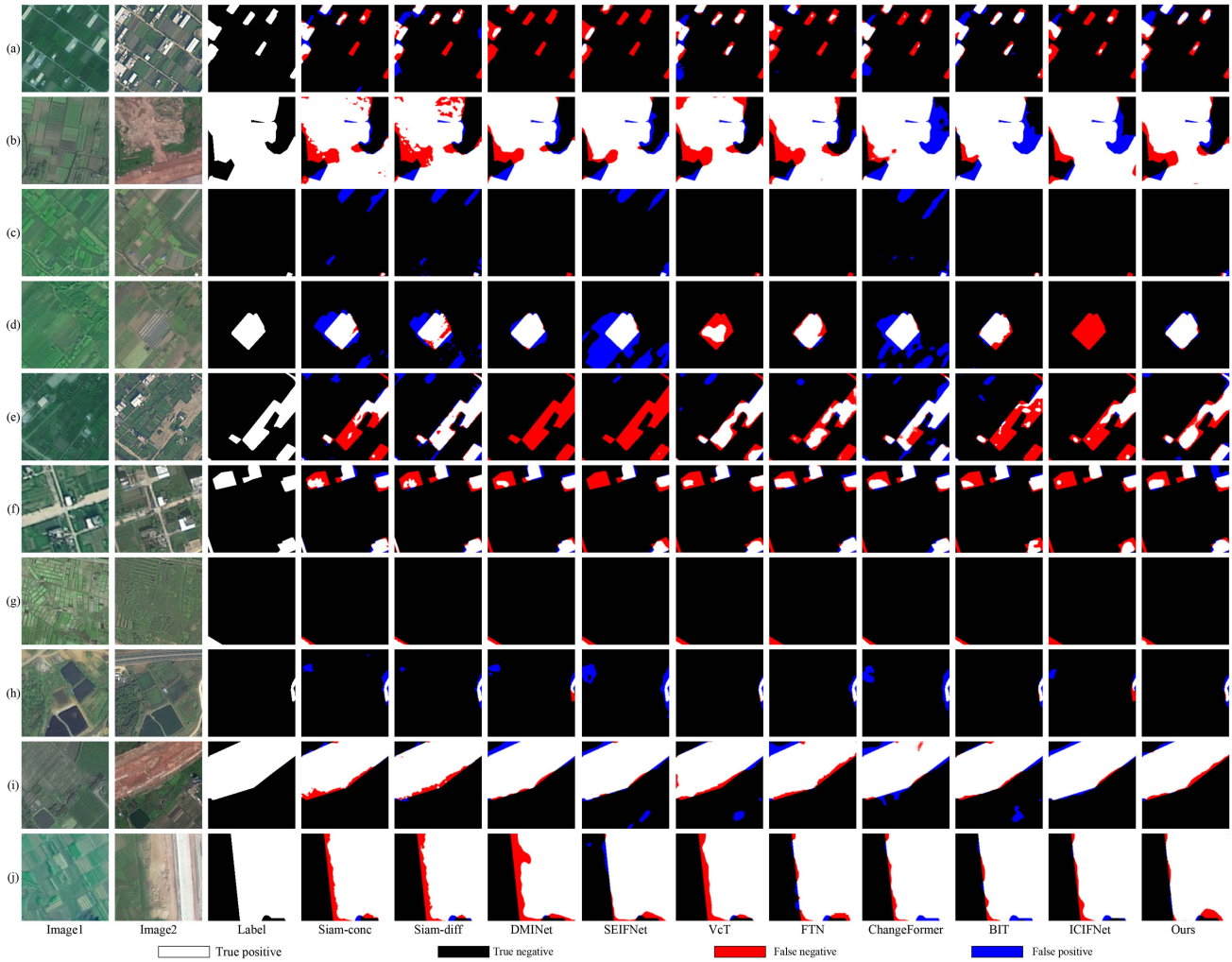


Fig. 5. Change detection results on CLCD.

recall, F1, IoU, and OA. Specifically, the optimal values of TFIFNet are 90.47%, 83.95%, 86.86%, 78.55%, and 96.55% respectively, which are 1.33%, 1.77%, 1.61%, 2.1%, and 0.39% higher than the suboptimal model FTN in various evaluation indicators. This may be because both the BFIM and the JointAtt effectively correct the localization of the changed regions, thus, improving the detection confidence of these regions. In addition, as the samples in CLCD dataset are farmland with relatively small intraclass differences, TFIFNet mines the potential intraclass relationships based on feature interaction, thus leading to the outstanding results. Although both TFIFNet and FTN prioritize the detection of changed areas, FTN adopts a fusion strategy that fuses multilevel visual features and ignores the impact of the semantic gap. TFIFNet adopts an adjacent layer feature fusion approach, which effectively mitigates the influence of the semantic gap on the network's performance. For other comparison methods, CLCD is a challenging dataset because its sample time span is wide, and the frequent seasonal changes and illumination variations lead to the existence of many spurious variables.

2) *Visualization Analysis on CLCD:* For the CLCD dataset, cropland with small intra-class differences is the main cause of changes. Since TFIFNet uses feature interactions twice to better mine the intraclass differences between samples, it can better divide the boundary of the changed region. For example, in Fig. 5(a), when there are many changed areas and the scale is generally small, TFIFNet can achieve better positioning and boundary division, while other SOTA models generally have the problems of boundary confusion and wrong positioning. When the change area is relatively large, as shown in Fig. 5(b) and (i), black holes or sawteeth generally exist in the SOTA model, while TFIFNet can identify relatively accurate boundaries. From the visualization results, we can see that TFIFNet can detect changes more effectively.

### F. Results and Analysis on SYSU-CD

1) *Quantitative Analysis on SYSU-CD:* The experimental results on the SYSU-CD dataset are shown in Table II. As can be seen from Table II, the performance of the proposed TFIFNet in precision, recall, F1 score, IoU, and

TABLE II
COMPARISON RESULTS ON SYSU-CD DATASET

| Method | Pre | Rec | F1 | IoU | OA |
|---|---|---|---|---|---|
| FC-Siam-conc [29] | 81.24 | 83.63 | 82.31 | 70.91 | 86.75 |
| FC-Siam-diff [29] | 83.33 | 81.30 | 82.24 | 71.00 | 87.58 |
| DMINet [59] | 83.58 | 84.89 | 84.21 | 73.58 | 88.39 |
| SEIFNet [65] | 80.20 | 84.58 | 81.92 | 72.27 | 85.97 |
| VcT [46] | 79.48 | 83.71 | 81.14 | 69.22 | 85.37 |
| FTN [61] | 90.32 | 86.31 | 88.08 | 79.34 | 91.84 |
| ChangeFormer [51] | 81.49 | 82.19 | 81.83 | 70.34 | 86.76 |
| BIT [47] | 83.10 | 82.50 | 82.80 | 71.70 | 87.71 |
| ICIFNet [30] | 82.64 | 82.96 | 82.80 | 71.66 | 87.54 |
| TFIFNet | **90.87** | **87.08** | **88.77** | **80.39** | **92.29** |

The bold face values indicate the best performing values on a single metric.

OA is similar to that of the CLCD dataset, which shows obvious superiority in all SOTA networks. We attribute this superior performance to our innovative fusion strategy that regards the adjacent layer's features as additional spatial information. With this strategy, TFIFNet achieves the best values in precision, recall, and IoU value, reaching 90.87%, 87.08%, and 80.39%, respectively. It also achieves the best F1 score and OA, which are 88.77% and 80.39%, respectively. TFIFNet outperforms the second best model FTN by 0.55%, 0.77%, 0.69%, 1.05%, 0.45% on all evaluation metrics. We think that this may be attributed to the fact that SYSU-CD contains more regions of large-scale variation and needs to make full use of the bitemporal information, which is ignored by most of the SOTA models. In the process of comparing the SOTA networks, we find that models that obtain the difference features by refining the bitemporal features, such as DMINet, FTN, and BIT, will have better detection results. We think this phenomenon benefits from the approach of obtaining difference features through refined bitemporal features, which helps to further mine and extract useful information from images, capture more details and contextual information, thereby improving the performance of CD.

2) *Visualization Analysis on SYSU-CD:* Compared with the CLCD dataset, the SYSU-CD dataset contains more large-scale change regions. Regarding the small change areas depicted in Fig. 6(e), most SOTA models exhibit significant limitations in the localization of changed areas, resulting in a higher misdetection rate. However, our network TFIFNet demonstrates superior accuracy in pinpointing these subtle changes. Furthermore, in Fig. 6(c), (h), and (i), where the change regions cover a substantial area, numerous SOTA models fail to effectively recognize the changes. Notably in Fig. 6(i), these models erroneously identify the change as encompassing the entire image scope. By comparing these visualization results, we can clearly see that TFIFNet is able to get closer to the real situation and provide more accurate detection results.

### G. Ablation Analysis

To validate the efficacy of each component in the proposed network, we perform ablation experiments using Pure as a baseline. Pure is a network of TFIFNet that removes BFIM and TFFM. All ablation experiments are performed on the CLCD dataset, and the performance trend is similar on the SYSU-CD dataset.

TABLE III
ABLATION EXPERIMENTS OF TWO IMPORTANT MODULES ON THE CLCD DATASET

| Module | BFIM | TFFM | Pre | Rec | F1 | IoU | OA |
|---|---|---|---|---|---|---|---|
| Pure | × | × | 86.48 | 78.65 | 82.00 | 72.48 | 95.41 |
| TFIFNet | ✓ | × | 88.63 | 79.11 | 83.06 | 73.76 | 95.75 |
| TFIFNet | × | ✓ | 88.53 | 82.31 | 85.10 | 76.24 | 96.08 |
| TFIFNet | ✓ | ✓ | 90.47 | 83.95 | 86.86 | 78.55 | 96.55 |

TABLE IV
ABLATION ANALYSIS ON THE NUMBER OF BFIM IN CLCD DATASET

| Num | Pre | Rec | F1 | IoU | OA |
|---|---|---|---|---|---|
| 1 | 89.06 | 84.28 | 86.48 | 78.03 | 96.38 |
| 2 | 90.47 | 83.95 | 86.86 | 78.55 | 96.55 |
| 3 | 89.01 | 83.65 | 86.12 | 77.56 | 96.31 |
| 4 | 89.10 | 83.66 | 86.13 | 77.58 | 96.31 |

TABLE V
ABLATION ANALYSIS ON THE INSERTION POSITION OF BFIM ON CLCD DATASET

| 1 | 2 | 3 | 4 | Pre | Rec | F1 | IoU | OA |
|---|---|---|---|---|---|---|---|---|
| ✓ | ✓ | | | 88.13 | 83.17 | 85.44 | 76.68 | 96.11 |
| | ✓ | ✓ | | 89.76 | 83.95 | 86.57 | 78.16 | 96.44 |
| | | ✓ | ✓ | 90.47 | 83.95 | 86.86 | 78.55 | 96.55 |

1) *Effect of BFIM:* Feature interactions can help to learn similar background distributions between the two branches and achieve domain adaptation between the two branches to some extent. The BFIM exchanges background distribution weights through spatial attention between the two branches of the proposed TFIFNet to ensure the reliability of changed regions. Table III shows that the BFIM module can improve precision, recall, F1, IoU, and OA by 2.15%, 0.46%, 1.06%, 1.28%, and 0.34%, respectively. Table III also demonstrated that feature interaction can improve the network's representation performance for changed features.

Furthermore, we also explored the impact of the number and the location of BFIM, the experimental results are shown in Tables IV and V, respectively. From Table IV, we can see that when the number of BFIM is 2, the performance of the network is optimal, and this is also the reference we adopted in this article. When the number is 3 or 4, the network's performance is worse than when the number of modules is 2 or 1. However, when the number of BFIM is 1, the second best result is achieved. We think that it may be because when the number of BFIM is large, the model pays more attention to the local invariant background information and ignores the global information, which leads to the model being unable to better locate the overall background distribution. From Table V, we can see that the performance is worse when the module insertion position is shallower, which we think may be because the shallow features contain mostly detailed information such as texture, which is distributed against the background.

2) *Effect of TFFM:* TFFM can make up for the loss of spatial information in the forward propagation process of the

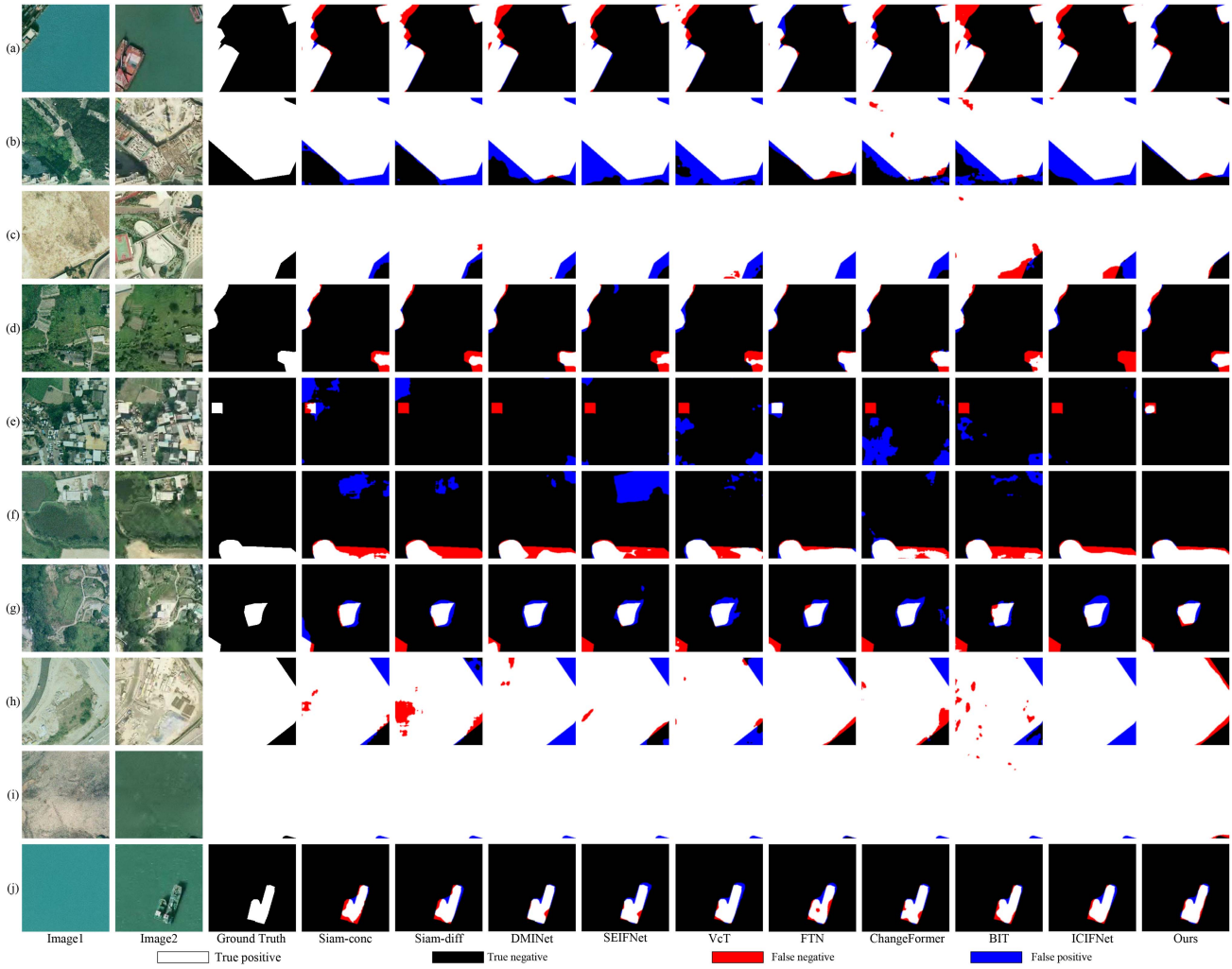


Fig. 6. Change detection results on SYSU-CD.

TABLE VI
ABLATION ANALYSIS OF THE FUSION STRATEGY AND COMPONENTS FOR THE TFFM ON THE CLCD DATASET

| Startegy | | Adj | JA | Pre | Rec | F1 | IoU | OA |
|---|---|---|---|---|---|---|---|---|
| Sum | Cas | | | | | | | |
| ✓ | | ✓ | ✓ | 88.96 | 83.43 | 85.94 | 77.32 | 96.27 |
| | ✓ | | ✓ | 87.82 | 82.66 | 85.01 | 76.12 | 96.01 |
| | ✓ | ✓ | | 90.30 | 83.67 | 86.63 | 78.24 | 96.49 |
| | ✓ | ✓ | ✓ | 90.47 | 83.95 | 86.86 | 78.55 | 96.55 |

network. Table III shows that triple feature fusion can improve the precision, recall, F1, IoU, and OA of the network by 2.05%, 3.66%, 3.10%, 3.76%, and 1.14%, respectively. The experimental results demonstrate that the triple feature fusion is the key of the proposed network TFIFNet.

In addition, to verify the effectiveness of the feature fusion strategy of the TFFM and the JointAtt component, we conducted experiments using the control variable method. The experimental results are shown in Table VI. When the simple addition method is used to achieve feature fusion and the adjacent layer features are used as additional spatial information, the overall performance of the network is poor. We believe that this is because direct addition can introduce noise, while concatenation can increase the nonlinear expression of features, thus making the features richer. When the adjacent layer's features are no longer used as additional spatial information, the performance of TFIFNet degrades significantly, with each indicator decreasing by 2.65%, 1.29%, 1.85%, 2.43%, and 0.54%, respectively. The experimental results in Table VI demonstrate that concatenating features along the channel dimension to achieve feature fusion and using adjacent layer features as additional spatial information are feasible and effective. When the JointAtt component is removed, the network's performance decreases by 0.17%, 0.28%, 0.23%, 0.31%, and 0.06%, respectively. This may be due to the fact that the BFIM module inserted into the deep layer of the encoder has already suppressed irrelevant variations to a large extent. Although it may be constrained by other components, the contribution of the JointAtt to the model's performance cannot be ignored.

### H. Model Complexity

Apart from numerical and visual results, to make better understand the efficiency of our network, we also report the

TABLE VII
COMPARISON OF MODEL EFFICIENCY

| Methods | Param(M) | FLOPs(G) | TrT(h) | InT(ms) | FPS |
|---|---|---|---|---|---|
| FC-Siam-conc | 1.55 | 8.16 | 0.93 | 2.12 | 470.96 |
| FC-Siam-diff | 1.35 | 7.24 | 0.88 | 2.09 | 478.12 |
| DMINet | 6.24 | 22.24 | 1.73 | 9.19 | 108.82 |
| SEIFNet | 27.90 | 12.82 | 0.78 | 8.93 | 111.93 |
| vcT | 3.50 | 16.29 | 2.6 | 44.76 | 22.34 |
| FTN | 164.45 | 88.82 | 1.97 | 43.29 | 23.10 |
| ChangeFormer | 29.75 | 32.43 | 4.9 | 26.76 | 37.38 |
| BIT | 3.50 | 16.28 | 1.13 | 10.61 | 94.26 |
| ICIFNet | 23.84 | 38.91 | 3.1 | 37.44 | 26.71 |
| TFIFNet | 168.60 | 103.94 | 1.48 | 51.81 | 19.30 |

parameter size (Param), floating-point-operations (FLOPs), train time (TrT), inference time (InT) as well as frames per second (FPS) of inference of our network and nine other SOTA methods, respectively. The results are shown in Table VII. From the table, we can see that CNN-based methods generally have a smaller number of Param, a generally lower FLOPs, and a shorter training and inference time. On the other hand, transformer-based and CNN-transformer-based methods generally have larger Param, higher FLOPs, and longer training and inference time, especially FTN and the proposed network TFIFNet, as they both use Swin transformer as encoder and decoder, while other methods generally use more traditional decoders. TFIFNet also introduced the attention mechanism in BFIM and JointAtt, resulting in the highest number of Param and FLOPs. However, the learning ability and the ability to capture complex features of the network are significantly improved, as it shows obvious advantages in five evaluation indicators. In practical applications, the deployment of complex models requires more resources, so how to make a tradeoff between performance and efficiency is also one of our future research directions.

## V. CONCLUSION

To take advantage of the proportion of invariant background in bitemporal images, we proposed a network called TFIFNet that is based on background distribution for CD. TFIFNet mainly consists of two components, namely BFIM and TFFM. The backbone is shared across both input images and used to extract bitemporal features. To mitigate the impact of irrelevant changes arising from registration errors, in the last two stages of the encoder, the BFIM is inserted to learn the distribution of the invariant background in bitemporal images. Second, to pay more attention to spatial detailed information, the TFFM, which integrates the JointAtt, utilizes adjacent bitemporal features as spatial detailed information that lost during the process of network training, and the temporal information contained in the bitemporal features is used to correct the difference features. Extensive experiments on two publicly available datasets, especially when the number of samples is limited, demonstrate that the TFIFNet based on the pretrained Swin Transformer can achieve state-of-the-art performance.

## DECLARATIONS

*Conflict of Interest:* The authors have no relevant financial or nonfinancial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

## ACKNOWLEDGMENT

*Author Contributions*: Dongen Guo: Methodology, investigation, writing-original draft, resource. Tao Zou: Software, conceptualization, validation, writing-original draft. Ying Xia: writing-review and editing, supervision, data Curation. Jiangfan Feng: writing-review & editing, supervision, funding acquisition.

## REFERENCES

[1] X. Song, Z. Hua, and J. Li, "Remote sensing image change detection transformer network based on dual-feature mixed attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416, doi: 10.1109/TGRS.2022.3209972.

[2] J. Yin et al., "Integrating remote sensing and geospatial Big Data for urban land use mapping: A review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102514, doi: 10.1016/j.jag.2021.102514.

[3] M. Kucharczyk and C. H. Hugenholtz, "Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. on. 112577, doi: 10.1016/j.rse.2021.112577.

[4] K. Rokni, A. Ahmad, A. Selamat, and S. Hazini, "Water feature extraction and change detection using multitemporal landsat imagery," *Remote Sens.*, vol. 6, no. 5, pp. 4173–4189, May 2014, doi: 10.3390/rs6054173.

[5] C.-F. Chen et al., "Multi-decadal mangrove forest change detection and prediction in Honduras, Central America, with landsat imagery and a Markov chain model," *Remote Sens.*, vol. 5, no. 12, pp. 6408–6426, Nov. 2013, doi: 10.3390/rs5126408.

[6] 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.

[7] J. Chen, P. Gong, C. He, R. Pu, and P. Shi, "Land-use/land-cover change detection using improved change-vector analysis," *Photogrammetric Eng. Remote Sens.*, vol. 69, no. 4, pp. 369–379, Apr. 2003, doi: 10.14358/PERS.69.4.369.

[8] 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, doi: 10.1109/TGRS.2008.916643.

[9] W. Feng, H. Sui, J. Tu, W. Huang, and K. Sun, "A novel change detection approach based on visual saliency and random forest from multi-temporal high-resolution remote-sensing images," *Int. J. Remote Sens.*, vol. 39, no. 22, pp. 7998–8021, May 2018, doi: 10.1080/01431161.2018.1479794.

[10] V. ARD and J. GR, "Five ways deep learning has transformed image analysis," *Nature*, vol. 609, pp. 864–866, Sep. 2022, Art. no. 7928, doi: 10.1038/d41586-022-02964-6.

[11] D. A. Kumar, M. Venkatanarayana, and V. Murthy, "Object-based image analysis," in *Encyclopedia of Mathematics Geoscience*, B. D. Sagar, Q. Cheng, J. McKinley, and F. Agterberg, Eds. Cham, Switzerland: Springer, Feb. 2023, pp. 1–5.

[12] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature Med.*, vol. 29, no. 9, pp. 2307–2316, Aug. 2023, doi: 10.1038/s41591-023-02504-3.

[13] X. Liu, J. Zeng, X. Wang, Z. Wang, and J. Su, "Exploring iterative dual domain adaptation for neural machine translation," *Knowl. Based Syst.*, vol. 283, no. 10, Jan. 2024, Art. no. 111182, doi: 10.1016/j.knosys.2023.111182.

[14] G. Deshmukh, O. Susladkar, D. Makwana, S. Mittal, and S. C. Teja R, "Textual alchemy: Coformer for scene text understanding," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2024, pp. 2931–2941.

[15] X. Huang, J. Zhang, D. Li, and P. Li, "Knowledge graph embedding based question answering," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 105–113.

[16] D. Guo, Y. Xia, L. Xu, W. Li, and X. Luo, "Remote sensing image super-resolution using cascade generative adversarial nets," *Neurocomputing*, vol. 443, no. 10, pp. 117–130, Jul. 2021, doi: 10.1016/j.neucom.2021.02.026.

[17] D. Guo, Y. Xia, and X. Luo, "GAN-based semisupervised scene classification of remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 12, pp. 2067–2071, Dec. 2021, doi: 10.1109/LGRS.2020.3014108.

[18] D. Guo, Y. Xia, and X. Luo, "Scene classification of remote sensing images based on saliency dual attention residual network," *IEEE Access*, vol. 8, pp. 6344–6357, 2020, doi: 10.1109/ACCESS.2019.2963769.

[19] A. S. Sagar, Y. Chen, Y. Xie, and H. S. Kim, "MSA R-CNN: A comprehensive approach to remote sensing object detection and scene understanding," *Expert Syst. Appl.*, vol. 241, no. 3, May 2024, doi: 10.1016/j.eswa.2023.122788.

[20] J. Hong, X. He, Z. Deng, and C. Yang, "Iou-aware feature fusion R-CNN for dense object detection," *Mach. Vis. Appl.*, vol. 35, no. 1, pp. 3–10, Nov. 2024, doi: 10.1007/s00138-023-01483-2.

[21] J. Hu, W. Zhong, M. Zhang, S. Kang, and O. Yan, "EiGAN: An explicitly and implicitly feature-aligned GAN for degraded image classification," *Pattern Recognit. Lett.*, vol. 178, no. 11, pp. 195–201, Feb. 2024, doi: 10.1016/j.patrec.2024.01.012.

[22] M. Zhang, L. Liu, Y. Jin, Z. Lei, Z. Wang, and L. Jiao, "Tree-shaped multiobjective evolutionary CNN for hyperspectral image classification," *Appl. Soft Comput.*, vol. 152, no. 4, Feb. 2024, Art. no. 111176, doi: 10.1016/j.asoc.2023.111176.

[23] C. Zheng, C. Hu, Y. Chen, and J. Li, "A self-learning-update CNN model for semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6004105, doi: 10.1109/LGRS.2023.3261402.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit.*, May 2015, pp. 3431–3440.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4690–4699.

[26] H. Ben Fredj, S. Bouguezzi, and C. Souani, "Face recognition in unconstrained environment with CNN," *Vis. Comput.*, vol. 37, no. 2, pp. 217–226, Jan. 2021, doi: 10.1007/s00371-020-01794-9.

[27] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017, doi: 10.1109/LGRS.2017.2738149.

[28] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2115–2118.

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

[30] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410213, doi: 10.1109/TGRS.2022.3168331.

[31] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 202, no. 10, pp. 599–609, Aug. 2023, doi: 10.1016/j.isprsjprs.2023.07.001.

[32] 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, doi: 10.3390/rs12101662.

[33] 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, 2022, Art. no. 8007805, doi: 10.1109/LGRS.2021.3056416.

[34] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019, doi: 10.1109/LGRS.2018.2869608.

[35] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, no. 3, Sep. 2021, Art. no. 102348, doi: 10.1016/j.jag.2021.102348.

[36] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021, doi: 10.1109/LGRS.2020.2988032.

[37] A. Vaswani et al., "Attention is all you need," in *Proc. 31th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.

[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 213–229.

[39] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput Vis. Pattern Recognit.*, Jun. 2019, pp. 2849–2858.

[40] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, May 2021, vol. 34, pp. 12077–12090.

[41] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput Vis. Pattern Recognit.*, Jun. 2021, pp. 6881–6890.

[42] D. Guo, Z. Wu, J. Feng, and T. Zou, "Multi-scale semantic enhancement network for object detection," *Sci. Rep.*, vol. 13, no. 1, May 2023, Art. no. 7178, doi: 10.1038/s41598-023-34277-7.

[43] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossVIT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 357–366.

[44] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Rep.*, Jan 2021.

[45] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10012–10022.

[46] B. Jiang et al., "VCT: Visual change transformer for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2005214, doi: 10.1109/TGRS.2023.3327139.

[47] 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.

[48] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610111, doi: 10.1109/TGRS.2023.3277496.

[49] K. Lu, X. Huang, R. Xia, P. Zhang, and J. Shen, "Cross attention is all you need: relational remote sensing change detection with transformer," *GIsci Remote Sens.*, vol. 61, 2024, doi: 10.1080/15481603.2024.2380126.

[50] N. Mungoli, "Adaptive feature fusion: Enhancing generalization in deep learning models," 2023, *arXiv:2304.03290*.

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

[52] 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, 2023, Art. no. 5615814, doi: 10.1109/TGRS.2023.3296383.

[53] Y. Ye, M. Wang, L. Zhou, G. Lei, J. Fan, and Y. Qin, "Adjacent-level feature cross-fusion with 3D CNN for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5618214, doi: 10.1109/TGRS.2023.3305499.

[54] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinsUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713, doi: 10.1109/TGRS.2022.3160007.

[55] X. Ma et al., "STNet: Spatial and temporal feature fusion network for change detection in remote sensing images," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2023, pp. 2195–2200.

[56] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, May 2022, doi: 10.1016/j.isprsjprs.2022.02.021.

[57] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615, doi: 10.1109/TGRS.2021.3130716.

[58] 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.

[59] 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, 2023, Art. no. 4401015, doi: 10.1109/TGRS.2023.3241257.

[60] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 2020, pp. 1580–1589.
[61] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proc. 16th Asian Conf. Comput. Vis.*, Feb. 2022, pp. 75–92.
[62] 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, doi: 10.1109/JSTARS.2022.3177235.
[63] 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, 2021, Art. no. 5604816, doi: 10.1109/TGRS.2021.3085870.
[64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[65] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609414, doi: 10.1109/TGRS.2024.3360516.

**Dongen Guo** received the B.S. degree in computer application and technology from Henan Normal University, Henan, China, in 2001, and the M.S. degree in computer application and technology from Wuhan University of Technology, Wuhan, China, in 2007, and the Ph.D. degree in computer science and technology from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2021.

He is currently a Professor and Vice Dean with the School of Computer and Software, Nanyang Institute of Technology, Nanyang, China. His current research interests include deep learning and computer vision.

**Tao Zou** received the B.S. degree in digit media technology from Southwest Petroleum University, Sichuan, China, in 2022. She is currently working toward the master's degree in computer technique from Chongqing University of Posts and Telecommunications, Chongqing, China.

Her research interest includes remote sensing image change detection.

**Ying Xia** (Member, IEEE) received the B.S. degree in computer science and technology from Chongqing University of Posts and Telecommunications, Chongqing, China, in 1993, and the M.S. degree in computer science and technology from Inha University, Incheon, South Korea, in 2001, and the Ph.D. degree in computer application technology from Southwest Jiaotong University, Chengdu, China, in 2012.

She is currently a Professor of computer science and technology with Chongqing University of Posts and Telecommunications. Her research interests include spatiotemporal Big Data, cross-media computing, and database systems.

Dr. Xia is a Member of the China Computer Federation.

**Jiangfan Feng** received the Ph.D. degree in cartography and geographical information system from Nanjing Normal University, Nanjing, China, in 2007.

He is currently a Professor with Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include GIS, artificial intelligence, computer vision, remote sensing, and VQA.