Exploring the Cross-Temporal Interaction: Feature Exchange and Enhancement for Remote Sensing Change Detection

Yikun Liu, Kuikui Wang, Mingsong Li, Yuwen Huang, and Gongping Yang

***Abstract*—Change detection (CD) aims to identify surface changes from bitemporal remote sensing (RS) images, which is a crucial and challenging topic in RS. In recent years, RS images CD has achieved significant advancements through the use of convolutional neural networks. However, existing deep learning-based CD methods still face some challenges, such as blurry boundaries, pseudochanges caused by fluctuations in imaging conditions, and complexity of change objects in the scene. In this study, we present a novel perspective for exploring the cross-temporal interaction in the feature fusion stage and propose a bitemporal feature exchange and enhancement network (ExNet). Specifically, we argue that the heterogeneity of bitemporal RS images leads to the failure of the CD model in some cases. Therefore, we attempt to achieve feature alignment in two aspects by feature exchange. On one hand, we exchange the statistical information of bitemporal features, facilitating the transfer of their style. On the other hand, bitemporal features are composed of content correlation embeddings and domain correlation embeddings (DCEs). We design a dynamic low-pass filter (DLF) to partially extract and exchange DCEs in bitemporal features to achieve the feature distribution alignment. Moreover, a frequency separation enhancement module is proposed, which transforms the fused features into the frequency domain and enhances the corresponding frequency representation. Comprehensive experimental results on three popular CD datasets demonstrate the effectiveness and efficiency of ExNet compared with state-of-the-art methods.**

***Index Terms*—Change detection (CD), dynamic low-pass filter (DLF), feature alignment, frequency domain enhancement.**

I. Introduction

REMOTE sensing (RS) change detection (CD) through image analysis techniques is utilized to compare and assess surface alterations between paired images taken at different periods in corresponding areas [1]. This method finds applications across different fields including urban planning [2], monitoring land use changes [3], and evaluating responses to disasters [4], [5].

Optical images have become a major source of data for CD due to their wide availability. Since these satellite sensors are capable of acquiring images with meter and submeter spatial resolution, fine spatial details of terrestrial objects can be studied [6]. However, optical very-high-resolution RS CD faces several challenges. First, there is heterogeneity in bitemporal RS images, which is due to the fact that bitemporal RS images are captured under diverse imaging conditions and at different time intervals, leading to variations in color or shape of ground objects caused by factors,such as seasonal changes, lighting conditions, and appearances. Second, high-resolution multitemporal RS images provide detailed information about ground objects, necessitating CD models to have strong feature recognition abilities. These challenges make it difficult to extract semantic features across temporal and spatial dimensions. Therefore, an effective CD model should exhibit robust cross-temporal interaction and semantic extraction capabilities.

So far, traditional CD methods can be generally divided into two primary categories: pixel-based CD and object-based CD [7]. Pixel-based CD techniques involve deriving the difference map directly from the spectral or texture characteristics of the bitemporal RS images. This method typically produces the change map by applying appropriate thresholds or clustering methods, such as change vector analysis [8], principal component analysis [9], and multivariate alteration detection [10]. To overcome the limitation of pixel-based CD methods in capturing spatial context details, object-based CD methods have been introduced. These techniques segment bitemporal RS images into regions based on spectral and spatial similarities, from which features are extracted to identify changes [11]. Object-based CD methods include algorithms, such as support vector machine [12], decision tree [13], and random forest [14]. Nevertheless, these traditional methods heavily rely on manually extracted features and often struggle to effectively represent high-level features, leading to reduced accuracy especially on high-resolution RS images.

In recent years, convolutional neural networks (CNNs) have been widely used in various RS tasks, such as scene classification [15], [16], [17], object detection [18], [19], semantic segmentation [20], and CD [21], [22], [23]. Several CNN-based

Manuscript received 8 April 2024; revised 15 May 2024, 30 May 2024, and 7 June 2024; accepted 10 June 2024. Date of publication 13 June 2024; date of current version 1 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176139, and in part by the Major Basic Research Project of Natural Science Foundation of Shandong Province under Grant ZR2021ZD15. *(Corresponding author: Gongping Yang.)*

Yikun Liu, Mingsong Li, and Gongping Yang are with the School of Software, Shandong University, Jinan 250100, China (e-mail: gpyang@sdu.edu.cn).

Kuikui Wang is with the School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China.

Yuwen Huang is with the School of Computer, Heze University, Heze 274015, China.

The source code will be accessible at: https://github.com/liuyk97/ExNet.

Digital Object Identifier 10.1109/JSTARS.2024.3413715

methods have been developed to improve the accuracy of CD. For example, one method utilized an edge-guided recurrent CNN that incorporated edge structure information to enhance CD [24]. Another technique involved generative adversarial training to create diverse bitemporal images, addressing issues related to data scarcity [25]. In addition, an unsupervised progressive learning framework was implemented to filter out incorrect change information in specific regions using a label selection filter [26]. Selective attention modules have been developed to explore the connection between semantics and pixel states, and these modules have been incorporated into metric learning-based architectures [27]. Moreover, a pair-to-video CD framework addressed CD as a video comprehension issue by generating a pseudotransition video [28]. In addition, generative adversarial networks have been employed to enhance super-resolution images and reduce resolution differences between bitemporal images [29].

Despite recent advancements, the ongoing diversity in bitemporal RS images poses a significant challenge, hindering CD methods from accurately pinpointing locations [30]. This issue stems from the heterogeneity of bitemporal features, resulting in variations in color and lighting of the same ground objects between different temporal RS images. While existing CD methods have made progress, their ability to harmonize domains in bitemporal RS images is still lacking. These methods often rely on complex model structures to enhance accuracy when merging bitemporal features [31], [32], [33], [34]. Whether bitemporal features are fused early or late in the CD network, two common fusion operators are utilized point-to-point differencing [28], [35] and channelwise concatenation [36], [37], [38]. Various feature fusion modules have been developed based on these fusion operations, including self-attentive structures [31], [33] and distance metrics [24], [39]. However, intricate model architectures may not fully tackle the diversity of bitemporal features. In addition, these feature fusion modules often prioritize the spatial aspect, overlooking valuable semantic information, such as the frequency domain. Consequently, biases exist in feature representation across different phases of RS images, significantly impacting the accuracy of CD models and resulting in cases of pseudochanges and missed changes [40].

To overcome these challenges, our strategy focuses on aligning domains for cross-temporal features and accurately capturing change features. With this in mind, we present a bitemporal feature exchange and enhancement network (ExNet) for RS image CD. Previous research has shown that exchanging partial features between bitemporal data can improve feature distribution similarity and aid in domain adaptation to some extent [41]. However, changer [41] employed tessellation masks for exchanging bitemporal features, which may jeopardize the integrity of the change object. To solve the above-mentioned problem and align the cross-temporal feature distribution, we implement feature exchange in the following two ways. First, recognizing that feature distributions are mainly shaped by their statistical properties [42], [43], we propose a recurrent exchange mechanism utilizing adaptive instance normalization (AdaIN) to exchange statistical information between bitemporal features. Second, we identify the existence of domain correlation embeddings (DCEs) and content correlation embeddings (CCEs) in bitemporal data. CCEs capture semantic details of objects that may undergo changes, as shown in Fig. 1 for buildings in the LEVIR-CD dataset, while DCEs represent background features with less semantic content but a significant impact on overall feature distribution. We theorize that differences in DCEs between bitemporal features contribute significantly to the variability observed in RS images. To address this, we propose a dynamic low-pass filter (DLF) to selectively extract and exchange DCEs between bitemporal features. In addition, acknowledging the importance of frequency domain information, we introduce a frequency separation enhancement module (FSEM) after the feature fusion step. FSEM comprises an amplitude enhancement branch and a phase enhancement branch to convert fused features into the frequency domain and enhance corresponding frequency representations individually.

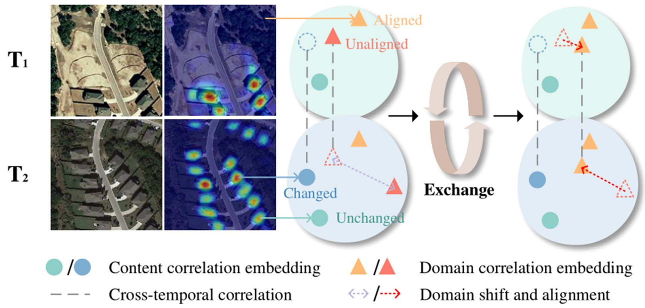
T₁
T₂
Aligned
Unaligned
Changed
Unchanged
Exchange
●/● Content correlation embedding
▲/▲ Domain correlation embedding
Cross-temporal correlation
Domain shift and alignment

Fig. 1. Motivation of the ExNet. The CCEs are associated with buildings, and the DCEs are associated with background, and the distance between the DCEs across temporal represents the domain shift between T_1 and T_2. The domain alignment is completed after the DCEs exchange.

The major contributions of this article are summarized as follows.

1) We address the heterogeneity of bitemporal RS images by introducing cross-temporal interaction, utilizing bitemporal features as content input and domain input, respectively. We align the channelwise mean and standard deviation of the content input to match those of the domain input.
2) We propose a novel perspective where DCEs and CCEs represent the semantics and distribution respectively in bitemporal features. Specifically, we design a DLF to partially extract and exchange DCEs. The resulting new bitemporal features theoretically possess similar feature distributions.
3) To further enhance fusion features, we introduce an FSEM, which boosts the amplitude and phase of frequency representation. This generates a comprehensive global information representation, enabling the extraction of more accurate change features.

The rest of this article is organized as follows. In Section II, the related work is introduced. The proposed ExNet is described in detail in Section III. The settings of all experiments are presented in Section IV. Section V presents and analyzes the experimental results. Finally, Section VI concludes this article.

II. Related Works

In this section, we briefly review the feature distribution alignment study and frequency domain study in the previous works.

A. Deep Learning-Based CD Methods

Currently, deep learning has advanced as an important technique for CD due to its ability to generate precise CD predictions. In contrast to traditional methods, deep learning-based methods can systematically extract spatial, spectral, and temporal features of bitemporal images in feature space. Consequently, deep learning-based methods attain superior CD accuracy, and several methods have been proposed to improve CD accuracy. Shi et al. [44] proposed a deeply supervised attention metric-based network, which learns change maps by means of deep metric learning. Zhang et al. [45] used a CNN to learn the deep features from RS images and used transfer learning to compose a two-channel network to generate a multiscale and multidepth feature difference map for CD. Recently, there has been a growing focus among researchers on the attention mechanism. Jiang et al. [46] proposed an efficient multidimensional attention-aggregation network, which keeps better feature aggregation while maintaining excellent differential attention ability. Li et al. [47] proposed a lightweight network, which identifies changes based on the features extracted by mobile networks via progressive feature aggregation and supervised attention. Zhang et al. [48] proposed a multiscale cascaded cross-attention hierarchical network, which uses a large kernel convolution formed by stacking small kernel convolutions combined with an efficient transformer as the backbone network to achieve local and global feature extraction and fusion.

Due to the strong representation ability of the transformer, the transformer-based models show comparable or even better performance as the convolutional counterparts in the CD task. Zhang et al. [49] introduced a U-shaped Swin-transformer-based network to improve the model's global exploration capabilities. This Siamese network employs a multilevel Swin-transformer block to construct an encoder and decoder framework, enhancing efficient CD. ChengeFormer [50] unified hierarchically structured transformer encoder with multilayer perception (MLP) decoder in a Siamese network architecture to efficiently render multiscale long-range details required for accurate CD. Zhang et al. [51] analyzed the relation changes in multitemporal images and proposed a cross-temporal difference attention to capturing these changes efficiently.

B. Feature Distribution Alignment Study for CD

Effectively addressing the heterogeneity of bitemporal RS images is a crucial issue in CD. Although deep learning has offered robust tools for enhancing CD models, many deep learning-based methods still rely on basic techniques, such as channel concatenation and pointwise differencing for interacting across different temporal states.

To enhance the interaction among diverse extracted features, researchers have suggested alternative fusion methods. For example, the channel-shuffle fusion strategy was proposed to exploit the strengths of different features by shuffling their channels [32]. Another method involved using a dual-feature mixed attention-based transformer network that combined fine and coarse features to handle misjudgments due to oversampling and ensure synchronized feature extraction and integration of target information [33]. Nonetheless, we argue that relying solely on complex model architectures may not adequately address the heterogeneity of bitemporal features.

Nonetheless, there are alternative methods that focus on aligning feature distributions by means of adversarial learning [29], [52], [53]. For instance, an innovative selective adversarial adaptation-based CD network was developed to capitalize on previously acquired knowledge from various domains and accomplish adaptation between multiple source domains and a target domain employing two domain discriminators [53]. Adversarial learning was also employed to create bitemporal images displaying varied changes in buildings [25]. Another strategy introduced two opposing Y-shaped networks to transform bitemporal images, preserving their inherent content characteristics while ensuring consistent style attributes [54]. Meanwhile, some researchers have worked on solving the problem of heterogeneity of bitemporal feature interactions in CD [55], [56]. Although these adversarial learning-based CD methods signify progress beyond mere concatenation and differencing, they predominantly hinge on comparative operations between two images. Chen et al. [57] proposed an unsupervised single-temporal CD framework based on intra- and interimage patch exchange. We argue that, for the effective alignment of feature distributions in bitemporal RS images, a simpler yet more robust method for intertemporal interaction should be investigated, which serves as the driving force behind our research.

C. Frequency Domain Study in Deep Learning Model

In the field of deep learning, analyzing signals in the frequency domain is vital for image processing tasks, such as image classification [58], dehazing [40], and pan-sharpening [59], [60]. In the realm of unsupervised domain adaptation for semantic segmentation, Fourier transform and its inverse are used for aligning domains [61]. In addition, a network combining frequency and spatial domains was proposed for comprehensive information utilization in CD [62]. Integration of wavelet transform into CNN models has been employed to resize feature maps, increase receptive fields, and reconstruct feature maps using inverse wavelet transform [63]. Some studies have employed dual-branch architectures to process amplitude and phase components separately, which led to the inspiration to improve these components individually.

Despite the sophistication of dual-branch frequency domain enhancement structures, their focus remains predominantly on local information in the frequency domain. We argue that to enhance frequency domain representation further, it is crucial to explore global information representation as well.

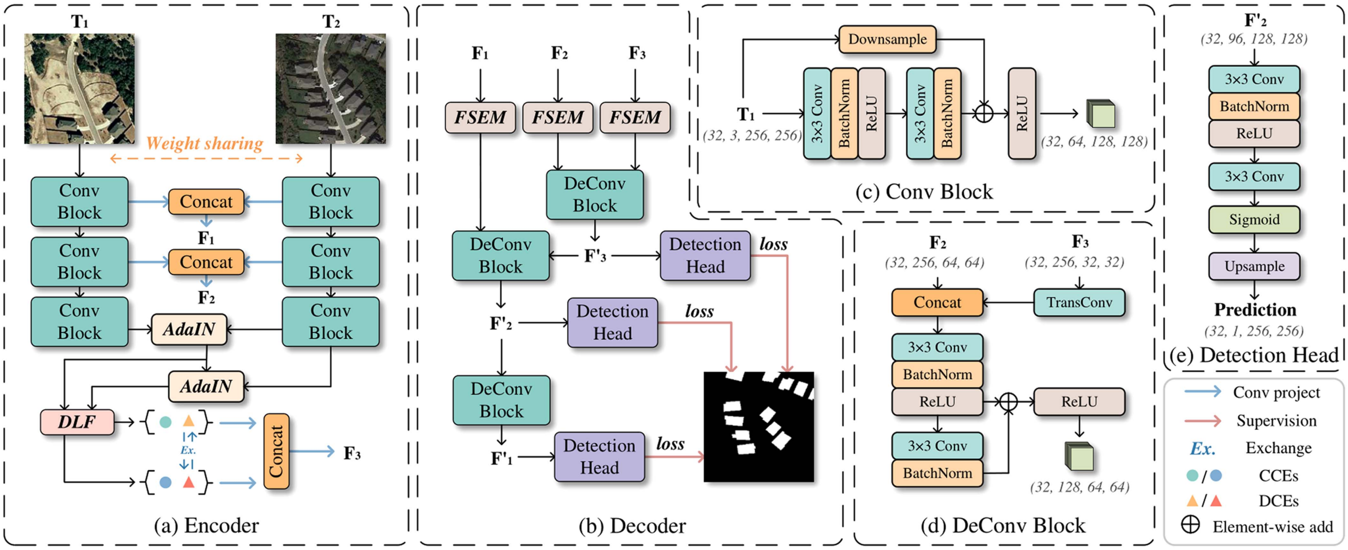

Fig. 2. Overview of the proposed ExNet.

III. Method

In this section, we propose a feature ExNet, which aims to address the heterogeneity of bitemporal RS images and extract accurate change features. We describe the overall flow of ExNet, including the exchange based cross-temporal interaction (ECI) and the FSEM. Finally, we discuss the loss function used in our work.

A. Overall Architecture

The ExNet architecture shown in Fig. 2 processes bitemporal RS images with three channels each to produce change prediction maps in a single channel. This architecture includes various components aimed at improving the accuracy of CD.

In the feature extraction phase, a dual-stream encoder with shared weight parameters captures features from input images, incorporating spatial and temporal information. Integration of information from both temporal phases is achieved in the initial layers by merging outputs of each convolution block.

To handle the inherent heterogeneity in bitemporal RS images, an exchange of statistical information and DCEs is conducted at the final convolution block outputs, aligning feature distributions between the images. This exchange, along with DCE data, reduces heterogeneity, thereby enhancing CD accuracy.

After feature fusion and distribution alignment, the fused features are processed in the FSEM with amplitude and phase enhancement branches. These branches operate on the frequency representation of fused features to boost change features through focus on frequency domain information.

Subsequently, a bottom-up decoder is then utilized for extracting change-related features, consisting of three deconvolution blocks that gradually upsample features while capturing finer details. This decoding step aids in reconstructing the change map to match the input images' spatial resolution.

Finally, the detection head refines the change map to attain a single-channel output. Deep supervision applied to the three detection heads enhances the model's discriminative ability for CD by incorporating intermediate supervision during training, thereby facilitating more effective learning.

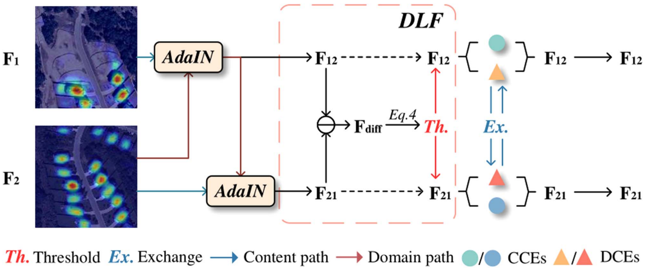

Fig. 3. Structure of the ECI.

B. Exchange Based Cross-Temporal Interaction

Our method aims to effectively deal with the diversity found in bitemporal RS images by incorporating two main parts: a recurrent exchange structure that facilitates the statistical information exchange of bitemporal features, and a DLF that enables the extraction and exchange of DCEs within these features. The core concept for the cross-temporal interaction, relying on the exchange mechanism, is depicted in Fig. 3.

The foundation of our method lies in the understanding that statistical information significantly influences feature distribution. Therefore, we utilize a style transfer method that utilizes AdaIN [42], an extension of instance normalization. AdaIN works by utilizing a content input represented as F_1 and a domain input denoted as F_2. Through aligning the channelwise mean and standard deviation of F_1 to those of F_2, AdaIN facilitates the transfer of style between bitemporal features

$$\mathrm{AdaIN}(F_1, F_2) = \sigma(F_2)\left(\frac{F_1 - \mu(F_1)}{\sigma(F_1)}\right) + \mu(F_2) \quad (1)$$

where $\sigma(x), \mu(x) \in \mathbb{R}^C$ are the mean and standard deviation, computed across batch size and spatial dimensions independently for each feature channel

$$\mu(x) = \frac{1}{NHW}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{nchw} \tag{2}$$

$$\sigma(x) = \sqrt{\frac{1}{NHW}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_{nchw} - \mu(x)\right)^2 + \epsilon} \tag{3}$$

where ϵ is a small constant. In this process, we utilize a scaling operation to standardize the input content F_1 using its standard deviation $\sigma(F_1)$ and mean $\mu(F_1)$, followed by a shift using the standard deviation $\sigma(F_2)$ and mean $\mu(F_2)$ of the domain input F_2, and vice versa. This method enables us to align the distributions of features during domain transfer by transferring channelwise mean and standard deviation statistics. Fig. 3 demonstrates the alternating utilization of F_1 as the content feature and F_2 as the domain feature to ensure distribution alignment of the dual-temporal features. In CD tasks, it is often the case that images in the later period contain more content, such as buildings, than images in the earlier period. So we first transfer the domain of F2 to F1, and then transfer the domain of F12 to F2, which can better preserve the content in F2.

Subsequently, we create the output feature F_{12} by merging the content embedding of F_1 with the domain embedding of F_2. Likewise, the output feature F_{21} is formed by combining the content embedding of F_2 with the domain embedding of F_1.

Recent studies have shown that exchanging bitemporal features partially during feature extraction can make the distribution between these features more alike, leading to a form of automatic domain adaptation between the bitemporal domains. Previous methods involved using tessellation masks to distinguish between exchange and nonexchange features, conducting feature exchange separately in the channel and spatial dimensions [41]. However, employing tessellation masks for exchanging bitemporal features is viewed as a drastic measure that could potentially jeopardize the integrity of the changed object. Hence, our goal is to investigate alternative exchange masks that can capture DCEs representing feature distribution without impacting the change content. Fig. 1 depicts the presence of DCEs and CCEs in bitemporal embedding. CCEs typically relate to objects undergoing changes, such as buildings in the LEVIR-CD dataset, which often contain more semantic information for CD. In contrast, DCEs usually denote background features with limited semantic information but are crucial for feature distribution. Therefore, exchanging partial DCEs of the bitemporal features facilitates aligning the domain distribution. Consequently, developing an exchange mask that accurately captures DCEs is vital. In this section, we present a proposal for a low-pass filter with a dynamic threshold to extract DCEs from the bitemporal features formed.

As illustrated in Fig. 3, CCEs generally have higher values due to their richer semantic content, while DCEs typically show lower values corresponding to the image background. Building upon this observation, a reasonable threshold can be utilized to differentiate between DCEs and CCEs. Moreover, it is recognized that with an increase in network training iterations, the network's reliability enhances, leading to a more precise differentiation between changing objects and backgrounds. To leverage this, we introduce a dynamic threshold for distinguishing CCEs and DCEs. Initially, we compute F_{diff} by carrying out point-to-point differencing between F_{12} and F_{21}. Subsequently, we determine the threshold T using (4), where features with low values are classified as DCEs and those with high values as CCEs. To promote cross-temporal interaction, we interchange partial DCEs among the bitemporal features, creating new bitemporal features F_{12} and F_{21} that theoretically exhibit similar feature distributions. The fundamental elements of this method are depicted in Fig. 3, and the threshold is defined as follows:

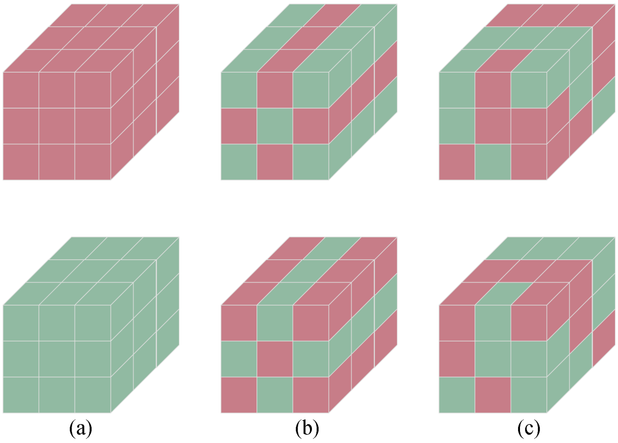

Fig. 4. Comparison of tessellation masks and our DLF. (a) Origin features. (b) Features exchanged with tessellation masks. (c) Features exchanged by our filter.

$$T = \mu(F_{\text{diff}}) - \left(1 - 2 \times \frac{\text{iter}_{\text{cur}}}{\text{iter}_{\text{max}}}\right) \times \sigma(F_{\text{diff}}) \tag{4}$$

where iter_{cur} and iter_{max} are the current and maximum number of iterations, respectively, the value of the features in F_{diff} larger than T are viewed as CCEs and vice versa as DCEs.

As illustrated in Fig. 4, the partial extraction and exchange of DCEs help maintain the integrity of modified objects and facilitate alignment of feature distribution through precise extraction of DCEs from bitemporal features.

C. Frequency Separation Enhancement Module

Frequency domain analysis plays a crucial and established role in image signal processing across various applications. It is valued for its stability, especially when compared to spatial domain processing. However, traditional methods for CD often overlook the need to restore distortions that arise in the frequency domain, a critical aspect for enhancing semantic representation. To address this gap, we present a new element called the FSEM, designed to efficiently capture and restore frequency representations.

In this section, we begin by revisiting the basic principles of Fourier transformations applied to images. Subsequently,

Algorithm 1: Exchange Based Cross-Temporal Interaction.

```
Input: $F_1, F_2 \in \mathbb{R}^{H \times W \times C}$.
for $iter_{\mathrm{cur}} = 1$ to $iter_{max}$ do
  Calculate the $\mu(F_1), \mu(F_2)$ and $\sigma(F_1), \sigma(F_2)$;
    Transfer the domain of $F_2$ to $F_1$, and generate the $F_{12}$;
    Calculate the $\mu(F_{12})$ and $\sigma(F_{12})$;
    Transfer the domain of $F_{12}$ to $F_2$, and generate the $F_{21}$;
    Calculate the $F_{\mathrm{diff}}$ by point-to-point differencing, and calculate the $\mu(F_{\mathrm{diff}})$ and $\sigma(F_{\mathrm{diff}})$;
    Generate the threshold $T$ with (4), bi-temporal features larger than $T$ are viewed as CCEs and vice versa as DCEs;
    Exchange the DCEs between bi-temporal features, and generate the new bi-temporal features $F_{12}$ and $F_{21}$.
  end for
Output: New bi-temporal features $F_{12}$ and $F_{21}$
```

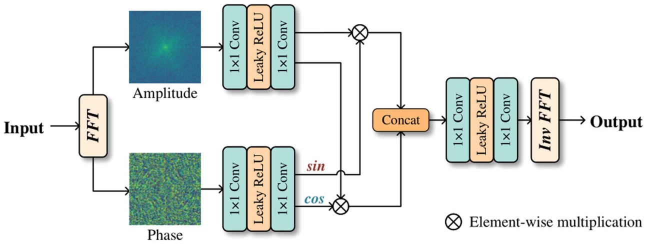

Fig. 5. Structure of the FSEM module.

we offer a detailed exploration of the structure of our FSEM, accompanied by thorough explanations.

Given a single channel image x with the Fourier transformer $F(x)$ converts to the Fourier space as a complex component, which is expressed as

$$\mathcal{F}(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)} \quad (5)$$

where $F^{-1}(x)$ defines the inverse Fourier transform accordingly. We separately apply Fourier transform to each channel in FSEM with the fast Fourier transform (FFT) [64].

In the Fourier space, each complex component $F(x)(u,v)$ can be represented by the amplitude component $\mathcal{A}(x)(u,v)$ and the phase component $\mathcal{P}(x)(u,v)$, which provides an intuitive analysis of the frequency components. $\mathcal{A}(x)(u,v)$ and $\mathcal{P}(x)(u,v)$ are expressed as

$$\begin{aligned} \mathcal{A}(x)(u,v) &= \sqrt{R^2(X(u,v)) + I^2(X(u,v))} \\ \mathcal{P}(x)(u,v) &= \arctan\left[\frac{I(X(u,v))}{R(X(u,v))}\right] \end{aligned} \quad (6)$$

where $R(x)$ and $I(x)$ represent the real and imaginary part of $F(x)$, respectively.

As depicted in Fig. 5, the first step of our method is to apply the FFT to convert the fusion features F into the amplitude and phase components. The Fourier transform of F can be mathematically expressed as follows:

$$\mathcal{A}(F), \mathcal{P}(F) = \mathcal{F}(F). \quad (7)$$

After extracting the amplitude and phase components using FFT, we establish two distinct operations to improve the depiction of $\mathcal{A}(F)$ and $\mathcal{P}(F)$. Each operation includes a 1×1 convolution and a subsequent application of the Leaky ReLU activation function. This process amplifies the distinguishing characteristics within both the amplitude and phase components effectively.

Next, we derive the actual and imaginary segments of the fusion feature by utilizing the improved amplitude and phase elements, represented as $\mathcal{R}(F)$ and $\mathcal{I}(F)$, correspondingly. This process enables us to encompass the details present in the amplitude and phase components distinctly. The formulation for this step is as follows:

$$\begin{aligned} R(F) &= \mathcal{A}(F) \times \mathrm{Cos}(\mathcal{P}(F)) \\ I(F) &= \mathcal{A}(F) \times \mathrm{Sin}(\mathcal{P}(F)). \end{aligned} \quad (8)$$

We combine the real and imaginary parts in a unique way to create a unified feature representation. This combined feature representation is then strengthened through the use of a 1×1 convolution layer and a Leaky ReLU activation function. This enhancement allows the model to capture broader global information that reflects the input features more accurately.

Finally, a global component separation step is carried out to extract the global information from the unified feature representation. Following this separation, an inverse fast Fourier transform is applied to convert the combined real and imaginary components back into the spatial domain. This procedure guarantees that the final output consists of spatially relevant features that encompass the improved global information. The process can be described as follows:

$$F_{\mathrm{out}} = F^{-1}(R(F), I(F)). \quad (9)$$

Expanding on Fourier theory principles, processing data in the Fourier space enables us to understand comprehensive global frequency representations in the frequency domain. In our methodology, we achieve the isolation and improvement of frequency representations using two distinct branch structures. These structural components are carefully designed to enhance semantic representation in the frequency domain. By treating amplitude and phase components separately and applying specific operations to enhance their representations, our method effectively boosts the semantic representation.

In summary, the frequency spectrum enhancement module (FSEM) facilitates individual enhancement of semantic representation for both amplitude and phase components. This comprehensive strategy results in the development of a representation that is enriched with global information.

D. Loss Function

In RS CD, a common challenge we face is the significant class imbalance between negative and positive pixels. This imbalance often hampers efficient optimization, leading to potential entrapment of the network in local minima of the loss function. To address this issue and enhance learning from complex scenes, we propose using a hybrid loss function that combines binary cross-entropy (BCE) loss and Dice loss. The hybrid loss function formulation is as follows:

$$L_{\text{hybrid}} = L_{\text{BCE}} + L_{\text{Dice}}. \tag{10}$$

To define the BCE loss and Dice loss, we view the predicted change map $\hat{Y}$ and the reference map Y as collections of pixels, denoted as $\hat{Y} = \{\hat{y}_i, i = 1, 2, \ldots N\}$ and $Y = \{y_i, i = 1, 2, \ldots N\}$. Here, $\hat{y}_i$ signifies the likelihood of change in the ith pixel, while y_i indicates the reference value in the ith pixel. In this context, a value of 0 signifies an unchanged pixel, and a value of 1 represents a changed pixel. The total count of pixels in the change map is designated as N. The combined objective function of both loss functions can be expressed as follows:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log_2 \hat{y}_i + (1 - y_i)\log_2(1 - \hat{y}_i). \tag{11}$$

The Dice loss is an effective option for tackling class imbalance issues and enhancing CD performance. It can be defined as follows:

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \hat{y}_i}. \tag{12}$$

Due to the utilization of deep supervision in the ExNet, we compute the hybrid loss for each of the predicted change maps as follows:

$$L_{\text{total}} = \sum_{s=1}^{S} L^{s}_{\text{hybrid}} \tag{13}$$

where S denotes the number of supervised stages, L^{s}_{hybrid} denotes the hybrid loss computed in the sth stage,

IV. Experiments Setting

A. Datasets

To assess the performance and robustness of our ExNet, we conducted comparative experiments using three challenging datasets. The first dataset is the season-varying change detection (SVCD) dataset [65]. The second dataset is the learning vision and remote sensing laboratory building change detection (LEVIR-CD) dataset [39]. The fourth dataset is the WHU-CD dataset [66]. These datasets provide diverse and complex scenes, allowing for a comprehensive evaluation of the effectiveness of our ExNet in CD tasks.

The SVCD dataset is specifically designed for general CD tasks. It consists of 15 998 pairs of bitemporal instances, where each pair demonstrates distinct appearances caused by seasonal factors. The images in the dataset have a size of 256 × 256 pixels. The dataset is divided into training, validation, and testing sets, with 10 000, 2998, and 3000 pairs of instances and corresponding labels, respectively.

The LEVIR-CD dataset comprises 637 image pairs, each with a size of 1024 × 1024 pixels and a spatial resolution of 0.5 m. To accommodate GPU memory limitations, the images are uniformly divided into 256 × 256 patches. As a result, the dataset contains 7120, 1024, and 2048 pairs of patches for training, validation, and testing, respectively.

The WHU-CD dataset is generated from one pair of aerial images captured from Christchurch, New Zealand. The spatial size of the dataset is 32 507 × 15 354, and the spatial resolution is 0.2 m. The dataset is also divided into small patches with size 256 × 256 for training, validation, and testing. The dataset is divided randomly into 5947/744/745 for training, validation, and testing, respectively.

B. Evaluation Metrics

In our experiments, we assess the models' performance using five metrics: Precision, recall, intersection over union (IoU), Kappa coefficient, and F1-score. These metrics offer a thorough evaluation of model performance, where higher values signify better performance.The formulations for these metrics are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{F}_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

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

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{18}$$

$$\text{P} = \frac{(\text{TP} + \text{FP})(\text{TP}+\text{FN})+(\text{FN} + \text{TN})(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})^2} \tag{19}$$

$$\text{Kappa} = \frac{\text{OA} - \text{P}}{1 - \text{P}}. \tag{20}$$

In the evaluation of a model's performance, TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) are used to represent specific counts. P denotes the theoretical probability of chance agreement between the ground truth and predictions.

Precision indicates the false alarm rate, showing the proportion of predicted positive instances that are truly positive. On the other hand, recall represents the missed alarm rate, indicating the proportion of actual positive instances correctly identified by the model.

Both precision and recall are crucial in evaluating a model, but there is often a tradeoff between them. Aiming for high precision usually results in lower recall, and vice versa. Achieving the right balance between these metrics depends on the task at hand and the desired compromise between reducing false positives and false negatives. It is vital to find the optimal balance based on the specific objectives of your application.

C. Implementation Details and Model Design

The model is built using the PyTorch framework and trained on an NVIDIA GeForce RTX 3090 GPU with 24 GB of video memory. Standard data augmentation techniques are used on input image patches to improve the model's generalization abilities. During training, a batch size of 32 is employed, and the initial learning rate is fixed at 5e-4 across all datasets, such as SVCD, LEVIR-CD, and WHU-CD.

The Adam optimizer is employed with a momentum of 0.9 and a weight decay of 0.0001. The optimizer's parameters β_1 and β_2 are, respectively, set to 0.9 and 0.99. To adapt the learning rate dynamically throughout training, the model adopts the polylearning scheme. The learning rate is adjusted based on the formula $(1 - (\text{iter}_{\text{cur}/}\text{iter}_{\max}))^{\text{power}} \times lr$, where the power value is 0.9.

For the LEVIR-CD datasets, the maximum number of iterations ($\text{iter}_{\max}$) is 40 000, while for the SVCD and WHU-CD dataset, $\text{iter}_{\max}$ is 160 000. These iterations are chosen to ensure adequate training for convergence and optimal model performance. These training configurations and hyperparameters are meticulously chosen to enhance the training process and maximize the model's performance on the respective datasets.

D. Comparative Methods

In this section, We make a comparison to several state-of-the-art methods, which are listed as follows.

1) FC-EF [36] utilizes an early fusion method, employing a UNet-based network architecture. It concatenates original bitemporal images as network input, processing them through a single-stream convolutional network for results.
2) FC-Siam-Diff [36] implements a feature-difference technique that extracts multilayer features of bitemporal images from Siamese networks for difference detection.
3) FC-Siam-Conc [36] follows a feature-concatenation method, extracting multilayer features of bitemporal images from Siamese networks for differential detection.
4) SNUNet [35] designs a densely connected deep SNN with strong feature extraction capabilities. It introduces an ensemble convolutional block attention module to refine features.
5) BIT [22] introduces a bitemporal image transformer within a deep feature differencing-based framework to model contexts in the spatial-temporal domain.
6) IFN [37] proposes a deeply supervised difference discrimination network for results. It fuses multilevel deep features with image difference features using attention modules for change map reconstruction.
7) Changer [41] introduces a new general results architecture incorporating two interaction strategies: Aggregation-distribution and "exchange." A comparison is made between its "ChangerEX" exchange architecture and our ExNet.
8) ChangeFormer [50] extracts the bitemporal features by cascaded transformer blocks, and the difference features were decoded by a lightweight MLP.

V. Results and Discussion

A. Comparison to State-of-the-Art Methods

To ensure a fair comparison, we reimplemented all comparative methods and replicated the results under the same experimental conditions. Optimal hyperparameters were selected for each method to maximize the F1 score on the validation subset.

Table I presents the quantitative results across the three datasets, with the highest values highlighted in bold for each column. For qualitative results, Figs. 6– 8 depict TP areas in white, FP areas in red, FN areas in green, and TN areas in black.

The changes observed in the SVCD dataset are more intricate compared to the other two datasets. While the LEVIR dataset and LEVIR+ dataset predominantly feature distinct buildings, such as houses, the SVCD dataset presents challenges as some altered regions resemble the background, leading to less accurate inferences.

1) Results on the SVCD Dataset: The quantitative results for precision, recall, and F1-score for all methods can be found in Table I. Notably, our proposed ExNet demonstrates superior performance, achieving precision and F1-score values of 97.96% and 97.23%, respectively. These results unequivocally showcase the exceptional performance of our CD method.

The performance outcomes on the SVCD dataset highlight the effectiveness of our method compared to results on the LEVIR-CD datasets. This difference is primarily due to the SVCD dataset's greater disparities in inherent characteristics, feature distributions, and a higher prevalence of change areas, making it a more challenging task. In contrast, methods such as FC-EF, FC-Siam-Diff, and FC-Siam-Conv demonstrate lower performance, indicating their limitations in the context of results. While the IFN method shows high recall (97.91%), it is prone to pseudochange near object boundaries. Similarly, changer exhibit subpar performance with F1-scores of 93.47%, 93.00%, 90.91%, and 90.93%, indicating their inefficiency in extracting distinctive features from diverse RS images. In conclusion, these results confirm that our ExNet excels in capturing frequency information and effectively identifying features in bitemporal RS images, resulting in superior CD performance compared to other methods.

Fig. 6 provides a visual comparison of the SVCD dataset, demonstrating the effectiveness of our method compared to other methods. Our technique excels in capturing subtle changes and accurately representing fine details of ground objects. Conversely, FC-EF, FC-Siam-Diff, and FC-Siam-Conv show limitations in extracting change features from diverse bitemporal RS images using Siamese network-based methods. To address these challenges, we have developed a feature distribution alignment method that includes transferring statistical information and exchanging DCEs.

The visualizations indicate that SNUNets and changer encounter difficulties in distinguishing shadows, often misidentifying building shadows as false changes. In contrast, our method surpasses others in detecting various irregularly shaped modifications in buildings. These visual results support the effectiveness of our proposed ExNet in capturing interactions across different time periods and accurately outlining object

TABLE I
QUANTITATIVE RESULTS ON THE SVCD, LEVIR-CD, AND WHU-CD DATASETS

Method	Params. (M)	Flops (G)	SVCD			LEVIR-CD			WHU-CD		
			Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
FC-EF	1.35	3.24	80.99	52.24	63.51	85.04	71.02	77.40	79.44	71.19	75.09
FC-Siam-Diff	1.35	4.38	87.51	44.37	58.89	88.98	84.71	86.79	49.27	79.29	60.77
FC-Siam-Conv	1.55	4.99	86.20	48.70	62.24	88.51	87.11	87.80	36.62	83.30	50.88
SNUNet	12.03	46.70	96.64	96.74	96.69	92.42	89.8	91.09	79.92	81.66	80.78
BIT	2.99	8.75	95.21	92.74	93.36	90.33	89.56	89.94	85.85	87.85	86.84
IFN	35.99	78.98	95.89	**97.91**	96.89	90.43	90.06	90.24	93.58	92.54	93.06
Changer	11.32	5.96	94.37	91.66	93.00	92.52	88.36	90.68	81.66	83.78	82.71
ChangeFormer	29.75	21.18	97.95	97.86	**97.90**	**93.04**	89.34	91.15	**95.38**	91.99	93.66
Ours	10.65	107.3	**97.96**	97.21	97.23	92.46	**90.84**	**91.64**	95.20	**92.56**	**94.02**

The highest score is marked in bold. All scores are described as percentages(%).

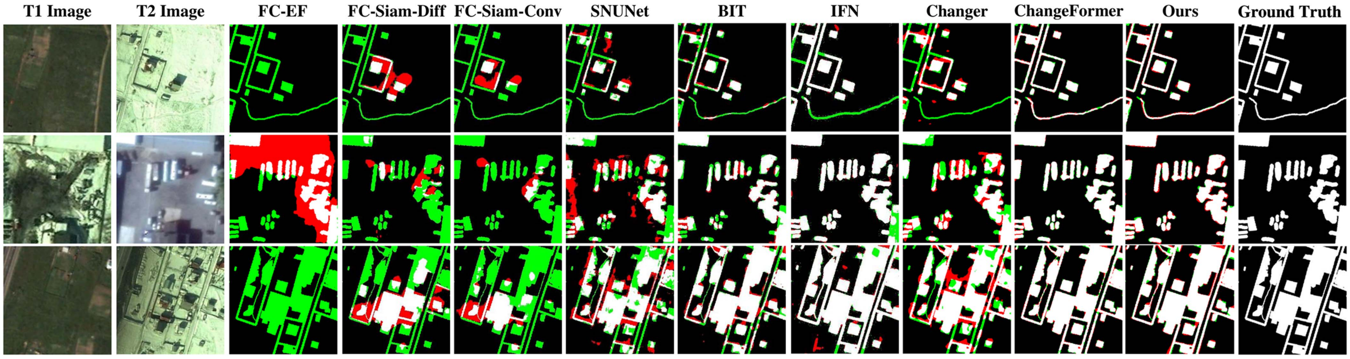

Fig. 6. Qualitative comparisons on the SVCD dataset. Different colors are used for a better view, i.e., white represents true positives, black represents true negatives, red represents false positives, and green represents false negatives. Where the less green and red, the better the performance represented.

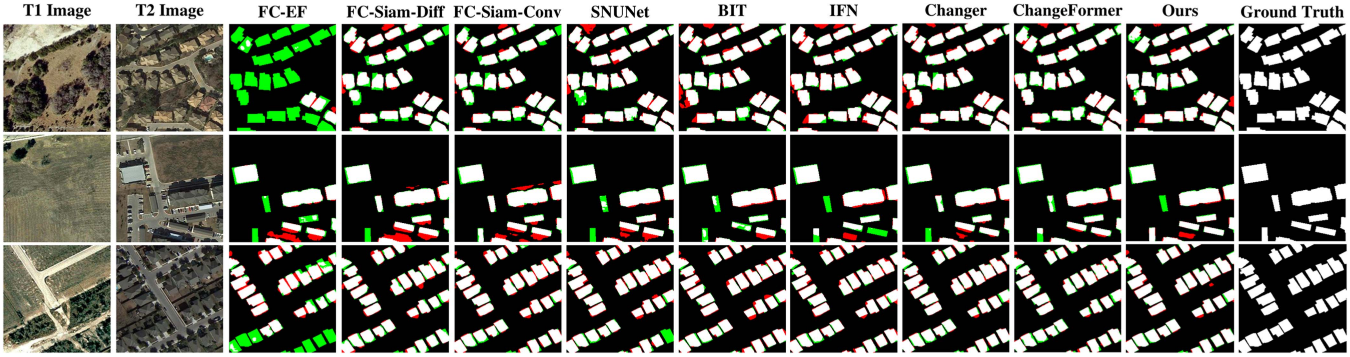

Fig. 7. Qualitative comparisons on the LEVIR-CD dataset. Different colors are used for a better view, i.e., white represents true positives, black represents true negatives, red represents false positives, and green represents false negatives. Where the less green and red, the better the performance represented.

boundaries. Furthermore, the underperformance of current methods on the SVCD dataset highlights the specific difficulties posed by the dataset's inherent characteristics and feature distributions. Our ExNet stands out for its ability to handle frequency information and promote effective feature interaction in bitemporal RS images, offering a robust solution to these challenges.

2) Results on the LEVIR-CD Dataset: As displayed in Table I, our new ExNet model outperforms existing methods significantly. It particularly shines in recall and F1-score, achieving 90.84% and 91.64%, respectively. In contrast, methods such as SNUNets face challenges due to the tradeoff between recall and precision, leading to the identification of false changes. Changer yields less than optimal results in precision and F1, with values of 93.05% and 90.68%, mainly due to its strategies involving spatial and channel exchanges. Noteworthy is that ExNet delivers more accurate predictions and notably decreases the occurrence of pseudochanges.

Fig. 7 presents a qualitative comparison using the LEVIR-CD dataset, showcasing the superior performance of our proposed

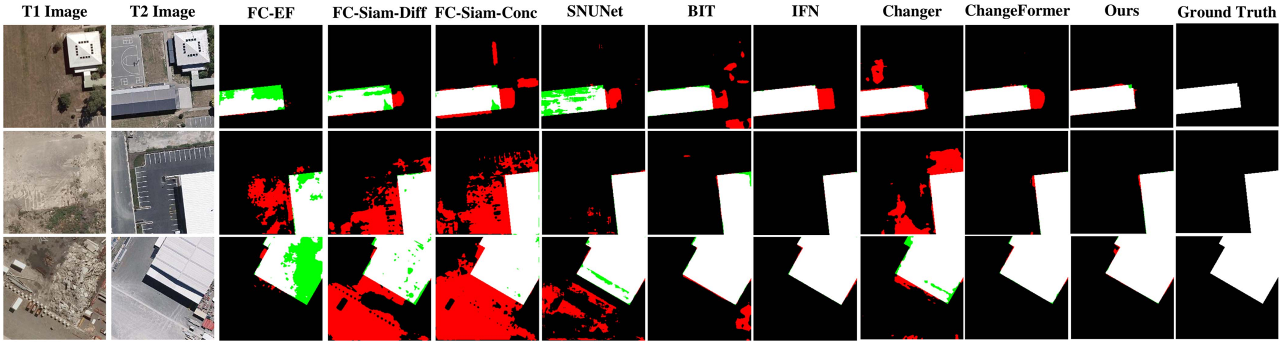

Fig. 8. Qualitative comparisons on the WHU-CD dataset. Different colors are used for a better view, i.e., white represents true positives, black represents true negatives, red represents false positives, and green represents false negatives. Where the less green and red, the better the performance represented.

method compared to existing ones. The challenge lies in accurately delineating pixels near edges where ground objects are small and have intricate features in RS images. The performance is further affected by the smaller size of the feature map in the multilayer feature extraction structure. While some methods aim to enhance edge detection by utilizing edge structures, they tend to create false changes near edges. In contrast, our method addresses this by incorporating feature alignment using two strategies: statistical information and DCEs through feature exchange. The visualizations in Fig. 7 illustrate that our method yields more refined results and achieves more precise object edge predictions. In addition, our method proves to be more resilient to shadow effects compared to others that often generate false changes due to shadow interference. In regions with densely distributed changes, our method excels in capturing details of minor changes effectively.

3) Results on the WHU-CD Dataset: The quantitative results for precision and F1-score for all methods are presented in Table I. Our ExNet model demonstrates superior performance with recall and F1 values of 92.56% and 94.02%, respectively. Although ChangeFormer exhibits the highest recall at 95.20%, its precision and F1 scores are comparatively low. Notably, ExNet strikes a better balance between precision and recall, resulting in higher overall performance and F1 scores compared to most benchmark algorithms.

Fig. 8 presents a qualitative comparison on the WHU-CD dataset, showcasing the superiority of our ExNet model over existing methods. Specifically, ExNet stands out in accurately detecting building edges, thanks to the FSEM that supplements missing frequency domain information. Conversely, FC-EF, FC-Siam-Diff, FC-Siam-Conv, and IFN show more pronounced false detections, primarily due to ineffective cross-temporal interaction. By incorporating ECI and extracting frequency domain information through FSEM, ExNet ensures consistency in objects and enhances CD along object boundaries. This enhanced performance highlights ExNet's effectiveness in managing complex cross-temporal interactions.

The consistent and robust performance of ExNet demonstrates its strong suitability for various CD tasks, including general-purpose and building change detection. In addition, our method shows resilience when dealing with datasets containing a considerable number of negative samples or when operating with limited sample sizes. This resilience can be credited to the efficient cross-temporal feature interaction and the accurate frequency domain information extraction by the FSEM. In conclusion, these results indicate that ExNet provides a dependable and potent solution for a wide range of CD tasks, enabling it to effectively address challenging real-world situations.

B. Complexity Analysis

In order to ensure a fair comparison of model efficiency, we assess the complexity of ExNet and other comparison methods based on the number of parameters (Params) and floating-point operations (FLOPs). Table I presents a detailed comparison of various networks in terms of Params and FLOPs. Our ExNet demonstrates superior performance while maintaining a reasonable number of parameters and FLOPs when compared to alternative methods. It is crucial to highlight that our model's improvement of frequency domain information significantly raises computational complexity, resulting in a higher FLOPs count. Conversely, simpler architectures such as FC-EF, FC-Siam-Diff, and FC-Siam-Conv have fewer Params and FLOPs. Nonetheless, they struggle to effectively address changes caused by scale variations, leading to subpar performance. In conclusion, our assessment suggests that ExNet strikes a favorable balance between model complexity and performance, positioning it as a practical choice for bitemporal RS image CD tasks. The results of the GPU inference time of the proposed method and the comparison methods are shown in Table II. The results in the table are obtained by calculating the average value of GPU inference time for all samples in the LEVIR-CD test dataset.

C. Ablation Study

The ablation study results are presented in Table III, which is intended to confirm how well each component of the ExNet performs, such as the ECI and FSEM. Two different sets of ablation experiments were carried out on the LEVIR-CD and WHU-CD datasets to provide a thorough demonstration of each module's effectiveness.

1) Ablation for ECI: The ECI's efficacy is evaluated through the gradual integration of AdaIN and DLF elements. The primary goal of ECI is to tackle the diversity present in bitemporal RS images by aligning feature distributions. The experimental

TABLE II
INFERENCE TIME OF THE PROPOSED METHOD AND THE COMPARISON METHODS

Methods	FC-EF	FC-Siam-Diff	FC-Siam-Conv	SNUNet	BIT	ChangeFormer	Ours
Time(ms)	7.17	9.61	10.10	15.72	17.23	13.18	43.44

TABLE III
ABLATION STUDY ON THE LEVIR-CD AND WHU-CD DATASET

Params. (M)	FLOPs (G)	ECI		FSEM			LEVIR-CD			WHU-CD		
		AdaIN	DLF	conv$_a$	conv$_p$	conv$_c$	Precision	Recall	F1	Precision	Recall	F1
4.49	51.39						92.40	90.20	91.29	94.13	89.44	91.75
5.67	52.60	✓					92.53	90.40	91.45	94.16	89.61	91.83
4.49	51.39		✓				92.60	90.01	91.29	93.13	89.44	91.25
5.67	52.60	✓	✓				92.92	89.98	91.43	94.24	90.59	92.38
5.80	54.05	✓	✓	✓			92.38	90.85	91.61	93.68	90.90	92.27
5.80	54.05	✓	✓		✓		92.98	89.93	91.43	94.85	90.56	92.66
10.39	104.41	✓	✓			✓	92.01	90.12	91.06	92.60	90.30	91.44
5.93	55.49	✓	✓	✓	✓		93.01	90.07	91.52	94.26	90.39	92.29
10.52	105.85	✓	✓		✓	✓	92.54	89.98	91.24	94.76	91.53	93.12
10.52	105.85	✓	✓	✓		✓	92.34	90.42	91.37	94.94	91.34	93.11
9.47	106.09			✓	✓	✓	92.89	89.88	91.36	93.93	90.13	92.33
10.65	107.30	✓	✓	✓	✓	✓	92.46	90.84	91.64	95.20	92.56	94.02

All scores are described as percentages(%).

TABLE IV
METRIC VALUES WITH COMPARED EXCHANGE MODULES ON THE LEVIR-CD AND WHU-CD DATASETS

Module	LEVIR-CD					WHU-CD				
	Precision	Recall	F1	IoU	Kappa	Precision	Recall	F1	IoU	Kappa
	92.40	90.20	91.29	83.97	90.83	94.13	89.44	91.75	84.73	91.41
Spex	91.45	90.76	91.10	83.66	90.63	92.99	89.90	91.42	84.20	91.07
Chex	92.45	90.40	91.41	84.19	90.96	92.81	90.59	91.69	84.65	91.36
Spex+Chex	92.24	89.87	91.04	83.55	90.57	93.70	90.36	92.00	85.19	91.68
AdaIN	92.53	90.40	91.45	84.25	91.00	94.16	89.61	91.83	84.89	91.50
DLF	92.60	90.01	91.29	83.98	90.83	93.13	89.44	91.25	83.89	90.90
AdaIN+DLF	92.92	89.98	91.43	84.21	90.98	94.24	90.59	92.38	85.82	92.05

All scores are described as percentages(%).

results unequivocally highlight the positive influence of ECI on classification accuracy. More specifically, there is a 0.14% enhancement in the F1-score for the LEVIR-CD dataset and a 0.63% increase for the WHU-CD dataset. These enhancements underscore the significant role of ECI in harmonizing feature distributions in bitemporal RS images, ultimately resulting in improved performance in CD tasks.

Importantly, it is significant to highlight that ECI manages to improve performance without substantially increasing the number of parameters and FLOPs. This highlights the advantage of ECI in improving ExNet's capability to match feature distributions in bitemporal RS images without a notable rise in computational complexity. In general, these results confirm the crucial contribution of ECI in enhancing ExNet's efficiency in handling the diversity of bitemporal RS images and improving feature distribution alignment, ultimately leading to enhanced CD performance.

To demonstrate the superiority of ECI over the exchange modules utilized in changer, we have integrated spatial exchange (Spex) and channel exchange (Chex) into the ExNet model. The results of our experiments, outlined in Table IV, unequivocally show that our feature exchange strategy surpasses the method employed in changer. This further supports our earlier claim that utilizing tessellation masks for exchanging bitemporal features may significantly impact the integrity of altered objects. In contrast, our method, which focuses on the selective extraction and exchange of DCEs, ensures the preservation of object integrity while aligning feature distribution. The precise extraction and exchange of DCEs from bitemporal features enable efficient feature exchange, resulting in enhanced performance compared to changer. These results highlight the effectiveness of our ECI and our method for addressing the heterogeneity of bitemporal RS images. Through the careful extraction and exchange of DCEs, our technique achieves feature distribution alignment without compromising the integrity of altered objects, ultimately enhancing CD performance.

In the proposed ExNet, the introduction of DCEs and CCEs in the bitemporal embedding aims to align feature distributions. CCEs carry more semantic information related to changing objects, whereas DCEs pertain to the background and play a crucial role in feature distribution. Experimental evaluations were carried out to assess the impact of different exchange times (1, 2, 3, and 4) on cross-temporal interaction. The results presented in Table V, suggest that increasing the number of exchanges does not enhance performance; instead, it leads to a decline in the F1 score. This decrease is attributed to inaccurate DCE exchanges,

TABLE V
Metric Values With Diverse Setting of Exchange Times on the LEVIR-CD and WHU-CD Datasets

Times	LEVIR-CD					WHU-CD				
	Precision	Recall	F1	IoU	Kappa	Precision	Recall	F1	IoU	Kappa
1	92.92	89.98	91.43	84.21	90.98	94.24	90.59	92.38	85.82	92.05
2	92.82	89.93	91.35	84.08	90.90	93.73	90.78	92.23	85.58	91.92
3	92.32	90.54	91.43	84.20	90.97	94.24	90.56	92.36	85.81	92.05
4	92.14	90.55	91.34	84.06	90.88	93.49	90.76	92.11	85.37	91.78

All scores are described as percentages(%).

TABLE VI
Metric Values With Diverse Setting of Threshold T on the LEVIR-CD and WHU-CD Datasets, σ, μ are the Mean and Standard Deviation, w is the Dynamic Weight

T	LEVIR-CD					WHU-CD				
	Precision	Recall		IoU	Kappa	Precision	Recall	F1	IoU	Kappa
μ	93.16	89.48	91.28	83.96	90.82	93.97	90.41	92.15	85.45	91.84
$\mu + \sigma$	92.59	90.17	91.36	84.09	90.90	93.88	90.88	92.36	85.80	92.05
$\mu - \sigma$	93.4	89.67	91.37	84.12	90.92	94.01	89.88	91.94	85.08	91.62
$\mu + w\sigma$	92.37	90.36	91.35	84.08	90.89	93.57	89.97	91.74	84.73	91.40
$\mu - w\sigma$	92.92	89.98	91.43	84.21	90.98	94.24	90.59	92.38	85.82	92.05

All scores are described as percentages(%).

which can confuse bitemporal features, affecting CCE representation and overall model performance. Interestingly, results show that exchanges at odd times (1 and 3) generally outperform those at even times (2 and 4), indicating that odd exchanges effectively align feature distributions, while even exchanges tend to restore original distributions. Consequently, based on these results, we opted for a single exchange in our experiments to balance feature alignment and preserve original representations. This choice allows for capturing DCEs accurately, thereby enhancing the model's performance in CD tasks.

In the ExNet model, the parameter T is crucial for defining the range of DCEs for feature distribution alignment. To enhance the extraction of DCEs and optimize bitemporal feature distribution alignment, we introduce a dynamic threshold strategy linked to the training iterations. To determine the best value for T, we experiment with various values of T and present the results in Table VI, where σ and μ denote the standard deviation and mean of the DCEs, respectively, and w represents the dynamic weight in (4). Our experiments show that the configuration specified in (4) yields the highest performance, indicating that the dynamic weight effectively adjusts to the boundaries between DCEs and CCEs. By dynamically adapting the threshold based on the mean and standard deviation of the DCEs, we can better identify the DCEs that require exchange, thereby enhancing feature distribution alignment. Consequently, in the ExNet model, we employ the dynamic threshold $T = \mu - w\sigma$ to determine the exchangeable DCEs. This dynamic threshold method improves the model's capacity to accurately select and exchange relevant DCEs, resulting in enhanced performance in CD tasks.

2) Ablation for FSEM: To enhance the frequency representations of bitemporal RS images effectively, we introduce a FSEM within the ExNet model. The FSEM improves the amplitude and phase components of fusion features independently and then boosts the combined real and imaginary parts. To assess the FSEM's effectiveness, we conduct ablation experiments by gradually eliminating the convolutional branches in the module. The results in Table III indicate that the FSEM enhances the model's performance on both the LEVIR-CD and WHU-CD datasets. Specifically, the F1 score on the LEVIR-CD dataset increases by 0.21%, while on the WHU-CD dataset, it rises by 1.64%. As shown in Table III, the convolutional branches (conv_a, conv_p, and conv_c) contribute to the performance enhancement with a slight rise in model complexity. Notably, conv_c, involving concatenation and doubling feature dimensions, significantly increases model complexity. Our rationale for integrating the FSEM is that frequency domain information offers better stability for CD tasks. Separating and enhancing global and semantic representations individually can lead to improved performance. By enabling the model to capture and restore frequency representations effectively, the FSEM enhances the ExNet model's discriminative power and overall performance in CD tasks.

In the conducted experiments, we applied the FSEM after each of the three stages of feature fusion in the ExNet model. The purpose of FSEM is to create global information representations and improve the semantic representations of the fusion features. To examine the effect of FSEM at different stages, we systematically removed FSEM from each stage and evaluated the model's performance. The results, detailed in Table VII, unequivocally illustrate that incorporating FSEM results in performance enhancements across all scenarios. This highlights the pivotal role of FSEM in generating global information representations and enhancing semantic representations at each fusion feature stage. The results underscore the importance of FSEM in capturing and restoring frequency representations in bitemporal RS images. By integrating FSEM at each stage, the ExNet model effectively utilizes frequency separation and enhancement benefits throughout the entire feature fusion process. This holistic

TABLE VII
METRIC VALUES WITH DIVERSE SETTING OF FSEM ATTACHED STAGES ON THE LEVIR-CD AND WHU-CD DATASETS

Stages	LEVIR-CD					WHU-CD				
	Precision	Recall	F1	IoU	Kappa	Precision	Recall	F1	IoU	Kappa
1	92.41	89.96	91.17	83.77	90.70	93.78	91.00	92.35	85.80	92.04
2	93.03	89.69	91.33	84.04	90.87	95.09	89.99	92.47	85.99	92.17
3	92.92	89.90	91.39	84.14	90.93	95.07	90.34	92.64	86.30	92.35
1, 2	92.82	90.09	91.44	84.23	90.99	95.59	90.90	93.19	87.25	92.91
2, 3	92.91	89.82	91.34	84.06	90.88	95.75	91.48	93.56	87.91	93.30
1, 2, 3	92.46	90.84	91.64	84.58	91.20	95.20	92.56	94.02	87.99	93.81

All scores are described as percentages(%).

TABLE VIII
METRIC VALUES WITH DIVERSE SETTING OF LOSSES ATTACHED STAGES ON THE LEVIR-CD AND WHU-CD DATASETS

Stages	LEVIR-CD					WHU-CD				
	Precision	Recall	F1	IoU	Kappa	Precision	Recall	F1	IoU	Kappa
3	92.39	89.97	91.16	83.76	90.70	95.87	89.30	92.47	85.99	92.17
2, 3	92.23	90.08	91.19	83.81	90.73	94.61	90.90	92.72	86.42	92.42
1, 2, 3	92.46	90.84	91.64	84.58	91.20	95.20	92.56	94.02	87.99	93.81

All scores are described as percentages(%).

TABLE IX
METRIC VALUES WITH DIVERSE SETTING OF LOSS FUNCTION ON THE LEVIR-CD AND WHU-CD DATASETS

Stages	LEVIR-CD					WHU-CD				
	Precision	Recall	F1	IoU	Kappa	Precision	Recall	F1	IoU	Kappa
BCE	93.03	89.49	91.23	83.87	90.76	94.89	91.19	93.01	86.93	92.72
Dice	92.54	90.42	91.47	84.28	91.02	94.85	91.71	93.25	87.35	92.98
BCE + Dice	92.46	90.84	91.64	84.58	91.20	95.20	92.56	94.02	87.99	93.81

All scores are described as percentages(%).

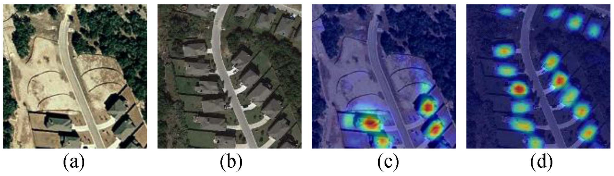
(a) (b) (c) (d)

Fig. 9. Illustration of DCEs and CCEs. Red denotes higher feature values and blue denotes low values.

inclusion of FSEM ensures that the model accurately captures and utilizes global and semantic information, ultimately leading to enhanced performance in CD tasks.

D. Discussion of DCEs and CCEs

In this study, we introduce an innovative method for handling cross-temporal feature interaction within the CD framework. Our analysis reveals that feature embeddings can be categorized into DCEs and CCEs. Fig. 9 illustrates that CCEs are linked to objects prone to change, such as buildings in the LEVIR-CD dataset, and encompass more semantically relevant information regarding alterations. Conversely, DCEs depict background characteristics, playing a significant role in the overall feature distribution while containing limited semantic details. Building on this result, our objective is to effectively extract and exchange DCEs to tackle the diversity present in RS images. To accomplish this, we devise a low-pass filter mechanism with an adaptive threshold. This thresholding procedure distinguishes features as DCEs if they possess low values, and as CCEs if they exhibit high values, enabling a clear separation of DCEs and CCEs within the feature embeddings.

To validate our method, we conducted multiple experiments and presented the results in Tables III, IV and V, IV. The results of these experiments offer substantial evidence in support of our theory, showcasing the efficacy of the proposed method. Through precise extraction and exchange of DCEs, our methodology effectively tackles the heterogeneity observed in RS images and enhances the performance of CD. The process of segmenting feature embeddings into DCEs and CCEs, followed by the extraction and interchange of DCEs, provides a practical and efficient means for cross-temporal feature interaction. This method not only retains the semantic information pertinent to CD but also effectively deals with the challenges arising from the diverse nature of RS images.

E. Discussion of Multilevel Loss

In this article, we present a hybrid loss function that merges BCE loss with Dice loss for CD. The hybrid loss aims to capture both pixel-level similarity and object-level overlap between predicted and actual change maps effectively. To assess the hybrid loss's efficacy, we perform experiments on models trained at various decoder stages, detailed in Table VIII. The findings reveal that employing the hybrid loss throughout all stages yields the most optimal performance. This implies that integrating the hybrid loss at each decoder stage enhances CD accuracy. Moreover, we compare models trained solely with BCE loss or Dice loss, outlined in Table IX. The results demonstrate that utilizing Dice loss outperforms using BCE loss alone. Furthermore, models trained exclusively with BCE or Dice loss exhibit inferior CD performance compared to those incorporating the hybrid loss. From these outcomes, we infer that the hybrid loss, combining BCE and Dice losses, proves beneficial for CD. By fusing pixel-level and object-level details, the hybrid loss boosts the model's capacity to precisely detect changes in input data. Hence, we opt to apply the hybrid loss after every decoder stage in our experiments.

VI. Conclusion

In this article, we investigate the cross-temporal feature interaction to address the heterogeneity present in bitemporal RS images. We posit that this heterogeneity can cause shortcomings in CD models. To tackle this issue, we suggest categorizing the bitemporal features into CCEs and DCEs to capture the semantics and distribution of the features, respectively. To align the features, we utilize two methods. First, we employ AdaIN to harmonize the style of the bitemporal features by aligning them based on their statistical characteristics. Second, we partially extract and exchange the DCEs to facilitate the alignment of feature distributions. Furthermore, we introduce an FSEM to enrich the frequency domain information within the combined features. The FSEM comprises an amplitude enhancement branch and a phase enhancement branch to enhance the amplitude and phase components independently, thereby refining the representation of frequency information in the features. The proposed ExNet undergoes evaluation on three prominent CD datasets and is benchmarked against cutting-edge methods. The experimental outcomes illustrate the effectiveness and efficiency of ExNet in achieving precise CD.

References

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.

[2] B. Demir, F. Bovolo, and L. Bruzzone, "Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 300–312, Jan. 2013.

[3] S. Jin, L. Yang, P. Danielson, C. Homer, J. Fry, and G. Xian, "A comprehensive change detection method for updating the national land cover database to circa 2011," *Remote Sens. Environ.*, vol. 132, pp. 159–175, May 2013.

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

[5] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size multitemporal images: Application to tsunami-damage assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1658–1670, Jun. 2007.

[6] D. Wen et al., "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 68–101, Dec. 2021.

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

[8] R. D. Johnson and E. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int. J. Remote Sens.*, vol. 19, no. 3, pp. 411–426, 1998.

[9] G. Byrne, P. Crapper, and K. Mayo, "Monitoring land-cover change by principal component analysis of multitemporal landsat data," *Remote Sens. Environ.*, vol. 10, no. 3, pp. 175–184, 1980.

[10] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998.

[11] E. A. Addink, F. M. B. Van Coillie, and S. M. De Jong, "Introduction to the GEOBIA 2010 special issue: From pixels to geographic objects in remote sensing image analysis," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 15, pp. 1–6, Apr. 2012.

[12] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS J. Photogramm. Remote Sens.*, vol. 61, no. 2, pp. 125–133, 2006.

[13] J. Im and J. R. Jensen, "A change detection model based on neighborhood correlation image analysis and decision tree classification," *Remote Sens. Environ.*, vol. 99, no. 3, pp. 326–340, 2005.

[14] K. J. Wessels et al., "Rapid land cover map updates using change detection and robust random forest classifiers," *Remote Sens.*, vol. 8, no. 11, Oct. 2016, Art. no. 888.

[15] X. Li, D. Shi, X. Diao, and H. Xu, "SCL-MLNet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5801112.

[16] M. Li, Y. Liu, G. Xue, Y. Huang, and G. Yang, "Exploring the relationship between center and neighborhoods: Central vector oriented self-similarity network for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1979–1993, Apr. 2023.

[17] X. Cheng et al., "Multi-view graph convolutional network with spectral component decompose for remote sensing images classification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2022.3227172.

[18] Y. Xi, L. Ji, W. Yang, X. Geng, and Y. Zhao, "Multitarget detection algorithms for multitemporal remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5400115.

[19] W. Xiong, Z. Xiong, Y. Cui, L. Huang, and R. Yang, "An interpretable fusion siamese network for multi-modality remote sensing ship image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2696–2712, Jun. 2023.

[20] Y. Liu, X. Kang, Y. Huang, K. Wang, and G. Yang, "Unsupervised domain adaptation semantic segmentation for remote-sensing images via covariance attention," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6513205.

[21] N. Bouhlel, V. Akbari, and S. Méric, "Change detection in multilook polarimetric SAR imagery with determinant ratio test statistic," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5200515.

[22] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416.

[23] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623811.

[24] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610613.

[25] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603216.

[26] Y. Zhou, X. Li, K. Chen, and S.-Y. Kung, "Progressive learning for unsupervised change detection on aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601413.

[27] M. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Edge neighborhood contrastive learning for building change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6001305.

[28] M. Lin, G. Yang, and H. Zhang, "Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images," *IEEE Trans. Image Process.*, vol. 32, pp. 57–71, 2023.

[29] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, “Super-resolution-based change detection network with stacked attention module for images with different resolutions,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403718.

[30] H. Zhang, M. Lin, G. Yang, and L. Zhang, “ESCNet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 28–42, Jan. 2023.

[31] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, “Deep multiscale siamese network with parallel convolutional structure and self-attention for change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406512.

[32] J. Lei, Y. Gu, W. Xie, Y. Li, and Q. Du, “Boundary extraction constrained siamese network for remote sensing image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621613.

[33] 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.

[34] 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.

[35] 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.

[36] R. Caye Daudt, B. Le Saux, and A. Boulch, “Fully convolutional Siamese networks for change detection,” in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[37] 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.

[38] A. Codegoni, G. Lombardi, and A. Ferrari, “TINYCD: A (not so) deep learning model for change detection,” *Neural Comput. Appl.*, vol. 35, no. 11, pp. 8471–8486, 2023.

[39] 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.

[40] H. Yu, N. Zheng, M. Zhou, J. Huang, Z. Xiao, and F. Zhao, “Frequency and spatial dual guidance for image dehazing,” in *Proc. Conf. Eur. Conf. Comput. Vis.*, 2022, pp. 181–198.

[41] 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.

[42] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1510–1519.

[43] S. Liu et al., “AdaAttN: Revisit attention mechanism in arbitrary neural style transfer,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6629–6638.

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

[45] M. Zhang and W. Shi, “A feature difference convolutional neural network-based change detection method,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020, doi: 10.1109/TGRS.2020.2981051.

[46] 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, 2023, Art. no. 4706118, doi: 10.1109/TGRS.2023.3328334.

[47] Z. Li et al., “Lightweight remote sensing change detection with progressive feature aggregation and supervised attention,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, 2023, Art. no. 5602812, doi: 10.1109/TGRS.2023.3241436.

[48] X. Zhang, L. Wang, and S. Cheng, “A multiscale cascaded cross-attention hierarchical network for change detection on bitemporal remote sensing images,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 62, 2024, Art. no. 2002216, doi: 10.1109/TGRS.2024.3361847.

[49] 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.

[50] W. G. C. Bandara and V. M. Patel, “A transformer-based Siamese network for change detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Kuala Lumpur, Malaysia, 2022, pp. 207–210, doi: 10.1109/IGARSS46834.2022.9883686.

[51] K. Zhang, X. Zhao, F. Zhang, L. Ding, J. Sun, and L. Bruzzone, “Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, 2023, Art. no. 5611615, doi: 10.1109/TGRS.2023.3281711.

[52] Z.-G. Liu, Z.-W. Zhang, Q. Pan, and L.-B. Ning, “Unsupervised change detection from heterogeneous data based on image translation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403413.

[53] M. Yang, L. Jiao, B. Hou, F. Liu, and S. Yang, “Selective adversarial adaptation-based cross-scene change detection framework in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2188–2203, Mar. 2021.

[54] B. Fang, G. Chen, G. Ouyang, J. Chen, R. Kou, and L. Wang, “Content-invariant dual learning for change detection in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603317.

[55] H. Chen, N. Yokoya, C. Wu, and B. Du, “Unsupervised multimodal change detection based on structural relationship graph representation learning,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5635318, doi: 10.1109/tgrs.2022.3229027.

[56] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, “Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 4700221, doi: 10.1109/TGRS.2021.3053571.

[57] H. Chen, J. Song, C. Wu, B. Du, and N. Yokoya, “Exchange means change: An unsupervised single-temporal change detection framework based on intra-and inter-image patch exchange,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 206, pp. 87–105, 2023.

[58] L. Gueguen, A. Sergeev, B. Kadlec, R. Liu, and J. Yosinski, “Faster neural networks straight from JPEG,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3937–3948.

[59] M. Zhou et al., “Spatial-frequency domain information integration for pan-sharpening,” in *Proc. Conf. Eur. Conf. Comput. Vis.*, 2022, pp. 274–291.

[60] M. Li, Y. Liu, T. Xiao, Y. Huang, and G. Yang, “Local-global transformer enhanced unfolding network for pan-sharpening,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2023, pp. 1071–1079.

[61] Y. Yang and S. Soatto, “FDA: Fourier domain adaptation for semantic segmentation,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4085–4095.

[62] Y. Zhou, Y. Feng, S. Huo, and X. Li, “Joint frequency-spatial domain network for remote sensing optical image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5627114.

[63] K. Xu, Z. Zhang, and F. Ren, “LAPRAN: A scalable Laplacian pyramid reconstructive adversarial network for flexible compressive sensing reconstruction,” in *Proc. Conf. Eur. Conf. Comput. Vis.*, 2018, pp. 485–500.

[64] H. J. Nussbaumer, *The Fast Fourier Transform*. Berlin, Germany: Springer, 1981.

[65] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, “Change detection in remote sensing images using conditional adversarial networks,” *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 422, pp. 565–571, May 2018.

[66] 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, doi: 10.1109/TGRS.2018.2858817.

Yikun Liu received the B.S. degree in mechatronic engineering from Huazhong Agriculture University, Wuhan, China, in 2019, and the M.S. degree in software engineering from Shandong University, Jinan, China, in 2022. He is currently working toward the Ph.D. degree with Shandong University.

His research interests include remote sensing image analysis and machine learning.

Kuikui Wang received the Ph.D. degree in software engineering from Shandong University, Jinan, China, in 2022.

She is currently a Lecturer with the School of Computer Science and Technology, Shandong Jianzhu University, Jinan. Her research interests include biometrics and pattern recognition.

Mingsong Li received the B.Eng. degree in software engineering from Shandong University, Jinan, China, in 2021, where he is currently working toward the M.S. degree in artificial intelligence.

His research interests include computer vision and machine learning.

Yuwen Huang received the Ph.D. degree in computer science and technology from Shandong University, Jinan, China, in 2021.

He is an Associate Professor with the School of Computer, Heze university, Heze, China. His research interests include biometrics and machine learning.

Gongping Yang received the bachelor's degree in computer and application, and the master's and Ph.D. degrees in computer software and theory from Shandong University, Jinan, China, in 1992, 2001, and 2007, respectively.

Since 2013, he has been a Professor with School of Software, Shandong University. His research interests include pattern recognition, computer vision, and remote sensing image analysis.

Dr. Yang is a Senior Member of CCF and CAAI, and also serve as the Machine Learning Technical Committee of CAAI and Artificial Intelligence and Pattern Recognition Technical Committee of CCF. During the past several years, he has served as a program committee Member/organization/program Chair of several conferences.