High-resolution remote sensing image change detection based on Fourier feature interaction and multi-scale perception.

CHEN, Y., FENG, S., ZHAO, C., SU, N., LI, W., TAO, R. and REN, J.

2024

High-Resolution Remote Sensing Image Change Detection Based on Fourier Feature Interaction and Multi-scale Perception

Yongqi Chen, Shou Feng, *Member, IEEE,* Chunhui Zhao, Nan Su, *Member, IEEE,* Wei Li, *Senior Member, IEEE*, Ran Tao, *Senior Member, IEEE*, and Jinchang Ren, *Senior Member, IEEE*

***Abstract*—As a significant means of Earth observation, change detection in high-resolution remote sensing images has received extensive attention. Nevertheless, the variability in imaging conditions introduces style discrepancies and a range of pseudo change regions between bi-temporal image pairs. Furthermore, changing objects possess diverse morphological representations, which makes accurately identifying change areas and delineating their boundaries within complex object distributions increasingly difficult. In response to the aforementioned challenges, we propose Fourier feature interaction and multi-scale perception (FIMP) model for effective change detection. To mitigate the impact of style discrepancies, FIMP employs the Fourier transform to adaptively filter bi-temporal features in the frequency domain, whilst mining the optimized bi-temporal features relevant to the change detection task. To enhance the ability to recognize multi-scale changing objects, FIMP aggregates and emphasizes the change areas with the introduced temporal change enhancement module (TCEM). By utilizing the U-fusion change perception module (UCPM) to perform multi-level bidirectional fusion of change features at different scales, FIMP can further enhance the ability to delineate complex semantic change boundaries. Experiments on three public datasets shows that our approach outperforms seven state-of-the-art methods.**

***Index Terms*—Change detection, high-resolution remote sensing image, Fourier feature interaction, multi-scale change feature.**

I. Introduction

CHANGE detection identifies surface alterations of the Earth by analyzing remote sensing images acquired at different times but from the same geographical location [1], [2]. These changes encompass a variety of phenomena, such as the construction of new buildings, the repair of road, the expansion of agriculture, the degradation of forests, environmental pollution, and so forth. Therefore, change detection has found wide applications in various fields, including urban planning [3], disaster assessment [4], and natural resource management [5]. With the rapid advancement of remote sensing technology, the accessibility of high-resolution remote sensing images is increasing, further driving the progress of change detection techniques.

Early traditional change detection methods can primarily be categorized into algebra-based approaches and transformation-based approaches [6]. Algebra-based methods obtain change detection results for each pixel through algebraic or statistical means, such as change vector analysis (CVA) [7] and spectral angle mapping (SAM) [8]. Transformation-based methods convert the original images into other feature spaces to distinguish changed and unchanged areas, such as principal component analysis (PCA) [9], independent component analysis (ICA) [10], and linear discriminant analysis (LDA) [11]. However, when faced with complex land cover distributions, traditional methods frequently exhibit limited detection performance. This limitation arises from their dependence on handcrafted features, which compromises robustness.

In recent years, with the widespread adoption of deep learning, an increasing number of researchers have begun to apply deep learning to change detection tasks [12]. Due to the powerful feature learning capabilities and hierarchical learning structures of convolutional neural networks (CNNs), change detection technology has been elevated to a new level. Fully convolutional early fusion (FC-EF) [13] adopts an early fusion method that directly concatenates bi-temporal images along the channel dimension as new input, performing end-to-end change detection tasks within a single stream framework. Nonetheless, it overlooks the consideration of temporal correlation, leading to an inability to accurately locate the changed areas. Compared to the single-stream architecture, the dual-stream framework utilizes siamese networks to extract multi-level abstract features from bi-temporal images [6]. It is more adept at learning and distinguishing the similarities or differences between bi-temporal instances. Consequently, subsequent researchers have shown a preference for adopting the dual-stream architecture to accomplish change detection tasks, such as Refs. [14], [15], [16], and [17]. With the emergence of attention mechanisms, many researchers have incorporated them into change detection models to achieve better feature representation. Attention mechanisms improve the efficiency and performance of change detection by computing weights for feature maps to capture key information relevant to the task, such as Refs. [18], [19], [20], [21], and [22]. However, due to the limitation of the fixed receptive field of convolutional kernels, the aforementioned methods struggle to capture the contextual information of the input images.

This work is supported by National Natural Science Foundation of China Grant 62471155 and 62371153, China Postdoctoral Science Foundation Grant 2023M740265, Open Fund of State Key Laboratory of Integrated Services Networks ISN25-05, the Fundamental Research Funds for the Central Universities under Grant 3072024LJ0804, 3072024XX0801, 3072024XX0805. (Corresponding author: Shou Feng.)

Yongqi Chen, Shou Feng, Chunhui Zhao and Nan Su are with the College of Information and Comminication Engineering, Harbin Engineering University, China, 150001; (e-mail:chenyq@hrbeu.edu.cn; fengshou@hrbeu.edu.cn; zhaochunhui@hrbeu.edu.cn; sunan08@hrbeu.edu.cn)

Shou Feng, Wei Li and Ran Tao are with School of Information and Electronics, Beijing Institute of Technology, Beijing, China; (e-mail:liw@bit.edu.cn; rantao@bit.edu.cn)

Jinchang Ren is with the National Subsea Centre, Robert Gordon University, AB21 0BH Aberdeen, U.K. (e-mail: jinchang.ren@ieee.org).

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

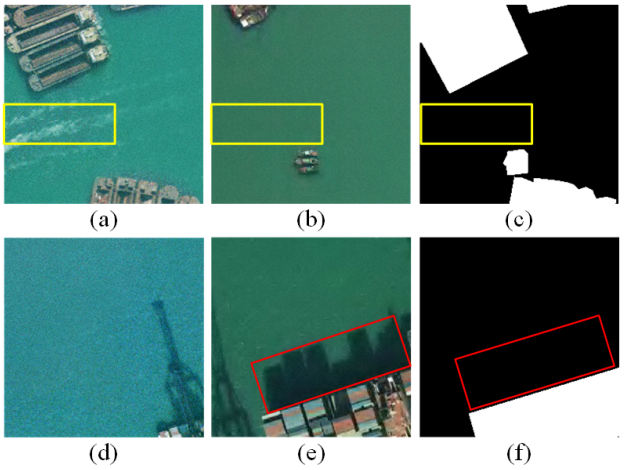


Fig. 1. The inconsistency in imaging conditions leads to stylistic discrepancies in bi-temporal images (yellow boxes), as well as building shadows caused by variations in solar elevation angles (red boxes). (a) and (d) are T1 image; (b) and (e) are T2 image; (c) and (f) are Ground truth.

To better capture long-range information and overcome the limitations of the fixed receptive field of convolutional kernels, many researchers have further introduced transformers into change detection models [23], [24].

However, the methods mentioned primarily concentrate on enhancing feature extraction from bi-temporal images and often overlook the unique characteristics of change detection tasks. Specifically, given the variability in imaging environments, maintaining consistent imaging conditions at the same geographical location over different times is exceedingly challenging for sensors. Therefore, inevitably, there are pseudo changes, which manifest as inconsistencies in the representation of areas that remain unchanged in bi-temporal images. These discrepancies can include shadows, brightness and stylistic discrepancies between the images. As highlighted in the yellow boxes in Fig. 1 (a)-(c), the color and brightness of these regions within the bi-temporal images are noticeably inconsistent. Additionally, there are waves or wakes present in non-changing areas. The red boxes in Fig. 1 (d)-(f) highlights the shadows cast by buildings due to variations in solar elevation angles during imaging. The presence of such pseudo changes can interfere with the effective extraction of change features by the model, leading to misinterpretations of change areas.

In view of the above problems, some researchers employ image translation to adjust the styles of bi-temporal images. This approach explicitly reduces the impact of pseudo changes resulting from inconsistencies in style between the images [25], [26], and [27]. However, the detection performance of the model largely depends on the quality of image translation. Therefore, some other researchers have started to adopt domain adaptation approach to implicitly mitigate the impact of pseudo changes [28] and [29]. Nevertheless, the aforementioned methods utilize multi-task learning or generative adversarial approaches to align the feature distributions of bi-temporal images, significantly increasing the complexity of model training.

Furthermore, in change detection tasks, the changing objects exhibit complex and diverse morphological representations. The diversity here is intuitively reflected in the varying sizes and irregular shapes of changing ground objects, which makes it more challenging to simultaneously recognize changing objects of different scales and accurately delineate their change boundaries.

To address the two major challenges, a change detection method for high-resolution remote sensing images based on Fourier feature interaction and multi-scale perception (FIMP) is proposed. Firstly, FIMP mitigates the inevitable style differences in change detection tasks from a frequency domain perspective. Considering that the change of a single element in the frequency domain will affect the global distribution of the original features [30]. FIMP filters the bi-temporal frequency domain features according to the task discrimination information contained in different frequency components to optimize the feature representation, thereby mitigating the influence of pseudo changes such as style discrepancies. Furthermore, to enhance the model's ability to accurately identify changing objects in complex scenarios, FIMP models the temporal correlation of bi-temporal images using the temporal change enhancement module (TCEM) to capture change areas. Subsequently, a U-fusion change perception module (UCPM) is introduced to aggregate the rich contextual information between multi-scale features [31]. This enhances model's ability to perceive the complex spatial forms of change objects, achieving precise localization of multi-scale change boundaries and change objects.

The main contributions can be summarized as follows:

1) To optimize the global representation of bi-temporal features with frequency domain feature, a Fourier feature interaction strategy is proposed, with a designed frequency domain feature interaction framework. This framework employs an adaptive frequency filtering module (AFFM) to adaptively weight different frequency components within bi-temporal features, which can not only optimize the global representation of features but also mine frequency components relevant to downstream tasks.
2) To fully model the temporal correlation of bi-temporal images and obtain change features, a temporal change enhancement module (TCEM) is proposed.TCEM captures more representative change information and highlights change regions through the full interaction with bi-temporal features. It provides preliminary guidance for the network to perceive multi-scale changing objects.
3) To aggregate the rich contextual information between multi-scale change features, a U-fusion change perception module (UCPM) is introduced. UCPM narrows the semantic gap between features of various scales through multi-layer bidirectional aggregation, enhancing the network's ability to perceive complex changing objects.

The remainder of this paper is organized as follows. Section II reviews some change detection methods based on deep learning and the Fourier transform of visual representation learning. Section III introduces the designed FIMP in detail. Section IV presents and analyzes the experimental results.

Finally, Section V concludes the work with summarized future directions.

II. Related Work

A. Deep Learning-Based Change Detection Methods

In recent years, change detection methods based on deep learning have shown great vitality. Daudt et al. [13] introduced the Siamese network into change detection, leveraging the shared-weight encoder to enhance the ability to detect changes. This dual-stream architecture facilitates the network's learning and differentiation between bi-temporal instances, which has led to its increasing adoption by researchers for change detection tasks. Zhao et al. [14] designed a novel dual-stream change detection framework using a semantic-guided strategy to overcome the interference issues present in bi-temporal feature fusion. Liu et al. [15], based on a dual-stream framework, achieved better detection results by aggregating spatial detail information and spectral difference information. Chen et al. [16] employed scale-invariant learning and local interaction to achieve change detection on continuous cross-resolution remote sensing images.

With the advent of attention mechanisms, many researchers have incorporated them into change detection models to enhance the focus on task-relevant information [19], [21]. Lv et al. [32] employed a spatial-spectral attention mechanism to capture more representative change features. Xu et al. [22] employed an attention pyramid and channel-cross attention mechanism to focus on crucial semantic information within channels and depth information. As a representative global self-attention method, Transformer has also achieved impressive results in change detection tasks [23]. Chen et al. [18] employed a Transformer to capture long-range information in bi-temporal images, improving the detection of change boundaries and small targets. Jiang et al. [24] utilized a transformer to learn a consistent representation of the background portions of bi-temporal images, enhancing the ability to recognize change areas.

Additionally, several methods have been proposed to enhance the performance of change detection tasks by incorporating advanced models [33]. Mei et al. [34] leveraged the powerful visual recognition capabilities of the segment anything model (SAM) to better extract and integrate contextual semantics, thereby improving the accuracy and robustness of semantic change detection. Dong et al. [35] explored the potential of multimodal data in the field of change detection by reconstructing contrastive language-image pretraining (CLIP) to extract bi-temporal features.

B. Fourier Transformation for Visual Representation Learning

In visual representation learning, Fourier transformation (FT) [36] is gaining attention as a valuable tool for frequency domain analysis. It extracts frequency information from images, enhancing models' understanding of global and local features. Some researchers have improved model generalization by processing images in the Fourier frequency domain, perturbing or enhancing components within specific frequency ranges [37]. FSDR [38] decomposes images into multiple frequency components, achieving domain generalization in semantic segmentation tasks by retaining domain-invariant components and randomizing domain-variant components. ALOFT [39] dynamically perturbs the low-frequency components of images, disrupting local textures while preserving global structural features, in order to reduce model overfitting to the source domain. Additionally, some researchers have explored incorporating FT into the network training process, combining frequency domain features with time domain features to enhance the model's learning capability [40]. FsaNet [41] designed a low-frequency self-attention module that better preserves edge information in semantic segmentation tasks. Liu et al. [30] fully exploited semantic information and fine-grained features by enhancing low-frequency features in the encoder and high-frequency features in the decoder.

Several researchers have integrated frequency domain analysis techniques into the domain of remote sensing image change detection [42]. Zheng et al. [43] designed a high-frequency-guided siamese network, utilizing a high-frequency attention module to enhance the network's focus on high-frequency information related to buildings. Tang et al. [44] employed wavelet transform to decompose features into different frequency components for separate interactions, completing object fine-grained change detection task.

The aforementioned methods each focus on different aspects of the frequency domain, but they do not utilize frequency domain information to address style discrepancies between bi-temporal images. In this paper, we transform the bi-temporal features into the frequency domain using FT and apply adaptive frequency filtering to facilitate interaction between the bi-temporal images, thereby mitigating the impact of style differences to some extent.

III. The Proposed Method

In this section, the overall framework of the proposed method is introduced. The Fourier feature interaction strategy employed by FIMP is detailed, followed by introductions to adaptive frequency filtering module (AFFM), temporal change enhancement module (TCEM) and U-fusion change perception module(UCPM). Finally, the loss functions used are discussed.

A. Overview

As shown in Fig. 2, the proposed FIMP mainly consists of a backbone, the Fourier feature interaction strategy (implemented by the AFFM), the TCEM, the UCPM, and a decoder. The first five layers of EfficientNet-b4, equipped with pre-trained weights, are used as the backbone network. Thanks to the effectiveness of the compound scaling strategy and the depth-wise separable convolutions, EfficientNet has a lower computational cost and competitive performance [45]. Therefore, it is highly suitable as a feature extractor in change detection tasks to obtain multi-scale feature maps from bi-temporal images.

Let a pair of bi-temporal images be $\{\mathbf{I}^1, \mathbf{I}^2\} \in \mathbb{R}^{H\times W\times 3}$, where H represents the height of the image, W represents the width of the image, and 3 represents the band of the image.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

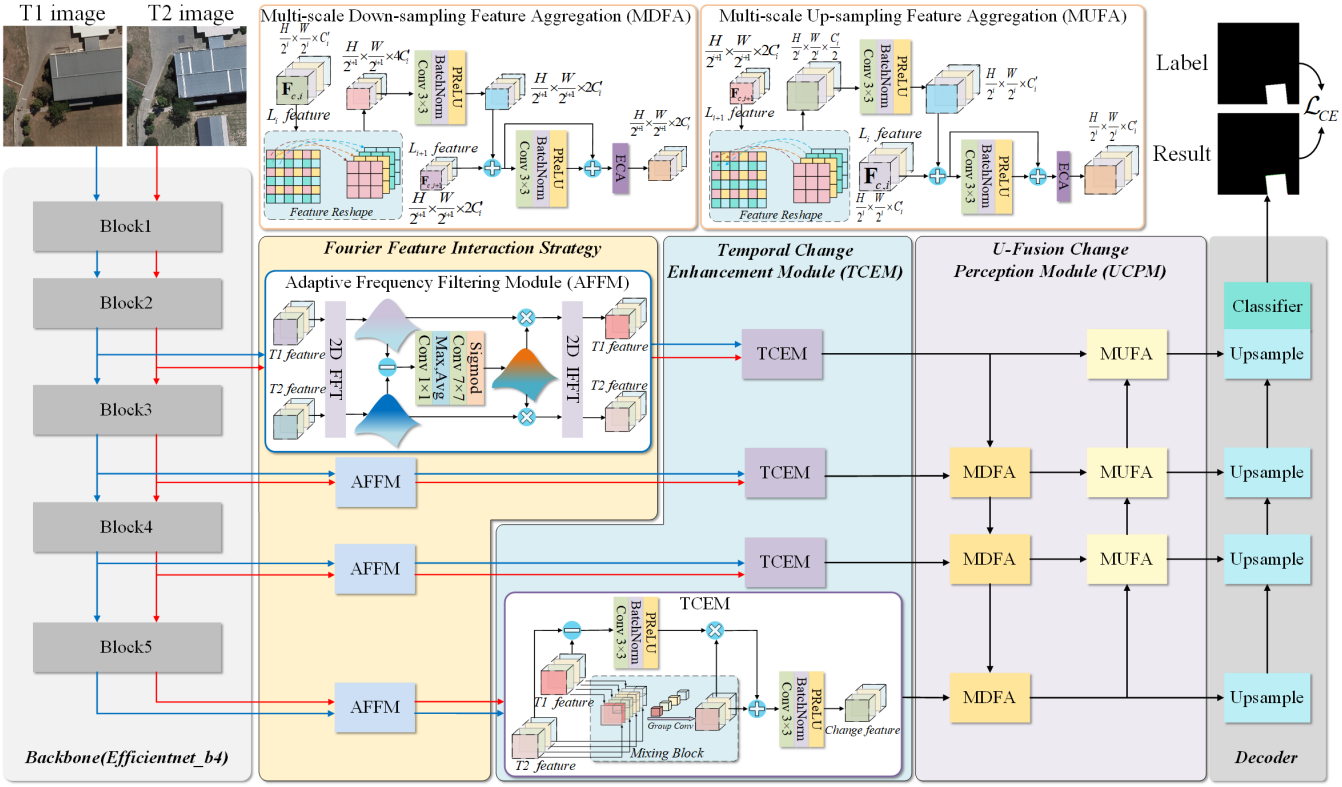


Fig. 2. The framework of FIMP, which is mainly composed of five parts: a pretrained backbone, Fourier feature interaction strategy implemented by the adaptive frequency filtering module (AFFM), temporal change enhancement module (TCEM), U-fusion change perception module(UCPM), and a decoder. In FIMP, block 1 to 5 represent the first five layers of EfficientNet-b4, which with pre-trained weights. The bi-temporal features are updated for global representation through AFFM. Multi-scale change features are obtained via TCEM, and the multi-scale feature semantics are aggregated using UCPM. Ultimately, the features are fed into the decoder to produce the prediction results.

After feeding the bi-temporal images into the backbone, multi-scale bi-temporal feature maps $\{\mathbf{F}_i^1, \mathbf{F}_i^2\} \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times C_i}$ are extracted, where $i \in \{0, 1, 2, 3, 4\}$ represents the ith layer of the backbone, $\frac{H}{2^i}$, $\frac{W}{2^i}$ and C_i represent the height, width and channel of the ith layer features, respectively. Subsequently, through the Fourier feature interaction strategy, the frequency features of bi-temporal images are optimized by using AFFM. By mining different frequency components in the feature map, the frequency components related to the change detection task are enhanced and the influence of redundant information is reduced. After adaptive frequency filtering, a TCEM is utilized. This module, operating in a manner similar to differential attention, fully interacts with bi-temporal features to capture change information across different scales of change objects.

To equip the model with the capability to capture the complex and diverse spatial morphological representations of change objects, UCPM is employed to aggregate feature maps across different scales. An architecture similar to U-Net is utilized to fully integrate multi-scale features, enabling the precise identification of multi-scale change targets while accurately locating complex change boundaries. Finally, the final change detection result map is obtained through layer-by-layer upsampling and a classifier.

B. Fourier Feature Interaction Strategy

For a given image, the low-frequency information is related to the slowly changing grayscale components, such as the image's color and brightness [38], [46]–[48]. The high-frequency information is related to the rapidly changing grayscale components, typically manifesting as the object's edge contours and speckle noise [30]. As shown in Fig. 3, after the fast Fourier transform (FFT) [36] of the bi-temporal images, the low-frequency component of the T1 image is replaced with the T2 image, and then the T2 image is reconstructed by inverse FFT. It can be observed that the color and brightness of the T2 image have changed and become similar to the T1

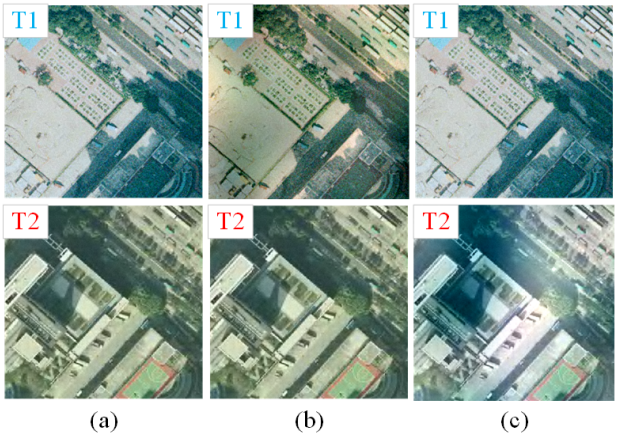


Fig. 3. After performing the Fourier transform on the bi-temporal images, replace their low-frequency components. The colors and brightness of the bi-temporal images are also replaced accordingly. (a) original bi-temporal images; (b) replacing the low-frequency components of the T1 image with those of the T2 image; (c) replacing the low-frequency components of the T2 image with those of the T1 image.

image, and vice versa. This further illustrates the correlation between frequency components and style representation. This characteristic aids in suppressing the impact of pseudo changes caused by style discrepancies between bi-temporal images through the use of frequency domain information.

During two-dimensional FFT [36] of an image, different frequency components in the image are decomposed into elements at different positions. This means that the entire image can be represented as a superposition of multiple frequencies, with each frequency corresponding to different structures and details in the image. Furthermore, in the frequency domain representation of an image, updating individual frequency domain elements affects the global representation of the entire image [49]. Based on this characteristic, a Fourier feature interaction strategy is proposed.

Firstly, the two-dimensional FFT [36] is applied to the spatial dimension of bi-temporal features $\{\mathbf{F}_i^1, \mathbf{F}_i^2\}$ after the backbone, transforming the features into the frequency domain. Since the features after FFT are in complex form, for ease of calculation, the real and imaginary parts are stacked along the channel dimension to obtain $\mathbf{F}_{FFT,i}^{1/2}$, which can be formulated as:

$$\begin{aligned}\mathbf{F}_{FFT,i}^{1/2}(u,v) &= \mathcal{F}\left[\mathbf{F}_i^{1/2}(h,w)\right] \\ &= \sum_{h=1}^{H-1}\sum_{w=1}^{W-1}\mathbf{F}_i^{1/2}(h,w)e^{-j2\pi\left(\frac{uh2^i}{H}+\frac{vw2^i}{W}\right)}\end{aligned} \tag{1}$$

where $\mathbf{F}_{FFT,i}^{1/2} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2C_i}$ is the feature obtained after applying the FFT to the ith layer's T1 temporal or T2 temporal feature.

After the FFT, the elements of the bi-temporal feature $\mathbf{F}_i^{1/2}$ are decomposed into frequency components at each position in $\mathbf{F}_{FFT,i}^{1/2}$. Therefore, a weighting matrix can be used to weight different frequency components to achieve a filtering effect, which can be formulated as:

$$\hat{\mathbf{F}}_{FFT,i}^{1/2} = \mathbf{F}_{FFT,i}^{1/2} \otimes \mathbf{M}_i \tag{2}$$

where $\hat{\mathbf{F}}_{FFT,i}^{1/2} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2C_i}$ is the feature after frequency filtering. $\mathbf{M}_i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 1}$ represents a weighting mask, and $\otimes$ denotes element-wise multiplication. It is worth noting that feature maps of different scales are filtered through weight masks corresponding to their resolutions. The method of generating these masks is described in Equation (6) to (7). This strategy optimizes the representation of bi-temporal features, thereby suppressing the impact of pseudo change information.

C. Adaptive Frequency Filtering Module (AFFM)

In FIMP, the Fourier feature interaction strategy is implemented by the adaptive frequency filtering module (AFFM). Specifically, after the backbone performs initial feature extraction on the input images, the features from the last four layers are used as the input for the AFFM. Through Equation (1), the bi-temporal features are transformed into the frequency domain. Subsequently, the differential frequency domain feature is obtained by subtraction, which can be formulated as:

$$\mathbf{D}_{FFT,i} = \mathbf{F}_{FFT,i}^1 - \mathbf{F}_{FFT,i}^2 \tag{3}$$

where the subtraction of bi-temporal features is carried out on an element-by-element basis. The frequency domain difference feature $\mathbf{D}_{FFT,i}$ can provide preliminary guidance for retrieving change information.

Since a simple subtraction is performed on the features from bi-temporal images, $\mathbf{D}_{FFT,i}$ contains redundant change information. If used directly to generate an attention map, it may have a negative impact. Therefore, a 1×1 convolution is first used to update different frequency components of the frequency domain features, thereby optimizing the feature representation. In addition, according to the linear properties of the Fourier transform : Fourier transform of the difference between the signal S1 and the signal S2 is equal to the difference between the Fourier transform of the two signals. That can be formulated as:

$$\mathcal{F}(S_1 - S_2) = \mathcal{F}(S_1) - \mathcal{F}(S_2) \tag{4}$$

where $\mathcal{F}(S_1)$ and $\mathcal{F}(S_2)$ respectively represent the Fourier Transforms of signal S_1 and S_2 signal.

In the frequency domain, a point does not represent a local area of the image, but rather information about a specific frequency throughout the entire image [50]. Therefore, the 1×1 convolution can also be considered a preliminary filtering of the original bi-temporal image difference features, representing an initial mining of change information. This process can be formulates as:

$$\hat{\mathbf{D}}_{FFT,i} = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}\left(\mathbf{D}_{FFT,i}\right)\right)\right) \tag{5}$$

where $\hat{\mathbf{D}}_{FFT,i}$ represents the updated difference features, σ refers to the Parametric Rectified Linear Unit (PReLU) used as a nonlinear activation function, and BN stands for Batch Normalization.

In order to mine the information carried by frequency components relevant to the change detection task during the frequency domain interaction process, an attention-based method is employed to adaptively weight different frequency components. In this context, the attention map is generated based on the differential feature $\hat{\mathbf{D}}_{FFT,i}$. This process can be formulated as:

$$\hat{\mathbf{F}}_{FFT,i}^{1/2} = \mathbf{F}_{FFT,i}^{1/2} \otimes \mathbf{A_i}\left(\hat{\mathbf{D}}_{FFT,i}\right) \tag{6}$$

where $\mathbf{A_i} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 1}$ represents a learnable attention mask. According to python 's broadcast mechanism, the generated attention mask and feature $\mathbf{F}_{FFT,i}^1$ perform channel-by-channel multiplication, thereby filtering out information irrelevant to the downstream task.

Inspired by the CBAM attention mechanism [51], the mask generation is based on the spatial attention mechanism in CBAM. This approach focuses on identifying the importance of various frequency components rather than distinguishing between individual channels. A shared weight mask across all channels ensures consistent attention to the same spatial locations, i.e., the frequency components, across the entire network. The differential features first undergo max pooling and average pooling along the channel dimension. Then, the frequency domain feature attention map is generated through

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

a 7×7 convolution followed by a Sigmoid function. The aforementioned process can be formulated as:

$$\begin{aligned}\mathbf{A_i} = \delta(\mathrm{Conv}_{7\times7}(\mathrm{Cat}[\mathrm{MaxPool}(\hat{\mathbf{D}}_{FFT,i}),\\ \mathrm{AvgPool}(\hat{\mathbf{D}}_{FFT,i})]))\end{aligned} \tag{7}$$

where δ represents the Sigmoid function, $\mathrm{MaxPool}(\cdot)$ refers to the max pooling layer, $\mathrm{AvgPool}(\cdot)$ refers to the average pooling layer, and $\mathrm{Cat}[\cdot]$ is the concatenation operation along the channel dimension.

According to Equation 3, the obtained attention map is used as weights to multiply with the bi-temporal frequency domain features to achieve filtering. Finally, the filtered bi-temporal frequency domain features $\hat{\mathbf{F}}^{1/2}_{FFT,i}$ are transformed back into the spatial domain through the Inverse Fast Fourier Transform (IFFT) to obtain $\tilde{\mathbf{F}}^{1/2}_i$ as follows:

$$\tilde{\mathbf{F}}^{1/2}_i = \mathcal{F}^{-1}\left(\hat{\mathbf{F}}^{1/2}_{FFT,i}\right) \tag{8}$$

where $\mathcal{F}^{-1}$ represents the IFT operation.

Because numerical changes of individual elements in the frequency domain can affect the global representation of the original features [49], such an approach can mitigate the impact of style differences. Moreover, since this is a differential frequency domain attention, it places more emphasis on preserving information relevant to the downstream change detection task during the filtering process.

D. Temporal Change Enhancement Module (TCEM)

For change detection tasks, thoroughly interacting and modeling the temporal correlations between bi-temporal images is imperative [52]. A common practice was to concatenate or subtract the features of bi-temporal images to model the temporal relationship between them, thereby identifying changing targets. However, when dealing with high-resolution remote sensing imagery with complex geographical distributions, such methods exhibit poor robustness and have limited ability to recognize changing targets. This is because such a method struggles to capture subtle spatial changes and the complex dynamic relationships between temporal phases.

To overcome the aforementioned shortcomings, a temporal change enhancement module (TCEM) is proposed. As shown in Fig. 2, the spatio-temporal dependency between two temporal phases is modeled by mixing the channel dimensions of the bi-temporal features. For the bi-temporal features after adaptive frequency filtering, feature interaction is first conducted through a mixing block. Specifically, for the bi-temporal features $\left\{\tilde{\mathbf{F}}^1_i, \tilde{\mathbf{F}}^2_i\right\} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C_i}$, they are first interlaced along the channel dimension, which can be formulated as:

$$\mathbf{F}_{mix,i} = \begin{cases}\tilde{\mathbf{F}}^1_{i,[k/2]} & if\ k\ is\ even\\ \tilde{\mathbf{F}}^2_{i,[k/2]} & if\ k\ is\ odd\end{cases} \tag{9}$$

where $\mathbf{F}_{mix,i} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2C_i}$ represents the mixed features, k is the index of the channel dimension of feature $\mathbf{F}_{mix,i}$. $\tilde{\mathbf{F}}_{i,[k/2]}$ and $\tilde{\mathbf{F}}^2_{i,[k/2]}$ denote the two-dimensional tensors of size $\frac{H}{2^i}\times\frac{W}{2^i}$ of the $[k/2]$-th channel of features $\tilde{\mathbf{F}}^1_i$ and $\tilde{\mathbf{F}}^2_i$, respectively. $[\cdot]$ denotes the downward rounding operation.

After the feature mixing, every adjacent pair of channels in $\mathbf{F}_{mix,i}$ represents the semantic information of the bi-temporal image, respectively. This characteristic of $\mathbf{F}_{mix,i}$ provides the possibility to model the spatio-temporal relationships between the bi-temporal images simultaneously. Therefore, the mixing block employs grouped convolution to interact the bi-temporal features under the condition of modeling spatio-temporal relationships. Specifically, for the feature $\mathbf{F}_{mix,i}$, group convolution is performed with a 3×3 convolution kernel, and the number of groups is C_i. During the convolution process, for a feature map with $2C_i$ channels, it is accordingly divided into C groups, meaning the dimension of each group of features is $\frac{H}{2^i}\times\frac{W}{2^i}\times 2$. Therefore, each group of convolution kernels can simultaneously perform spatial and temporal convolution on the bi-temporal features of the corresponding channels, fully exploiting the morphological representation of the geographical features and the temporal change relationships.The above process can be formulated as follows:

$$\hat{\mathbf{F}}_{mix,i} = \sigma\left(\mathrm{BN}\left(\mathrm{GroupConv}_{3\times3}\left(\mathbf{F}_{mix,i}\right)\right)\right) \tag{10}$$

where σ represents the PReLU activation function, BN stands for batch normalization layer, and $\mathrm{GroupConv}_{3\times3}$ is the grouped convolution layer with a kernel size of 3×3.

To enhance the change information contained in $\hat{\mathbf{F}}_{mix,i}$ and optimize the feature representation, the bi-temporal features $\left\{\tilde{\mathbf{F}}^1_i, \tilde{\mathbf{F}}^2_i\right\}$ are subtracted to highlight the change information. The differential change information is then conveyed into feature $\hat{\mathbf{F}}_{mix,i}$ in a manner similar to attention weights. This process can be formulated as:

$$\mathbf{D}_i = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times3}\left(\left|\tilde{\mathbf{F}}^1_i - \tilde{\mathbf{F}}^2_i\right|\right)\right)\right) \tag{11}$$

$$\mathbf{F}_{c,i} = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times3}\left(\hat{\mathbf{F}}_{mix,i}\otimes\mathbf{D}_i\oplus\hat{\mathbf{F}}_{mix,i}\right)\right)\right) \tag{12}$$

where $\mathbf{F}_{c,i} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C'_i}$ represents the output of the i-th TCEM, and $\oplus$ represents element-wise addition.

As illustrated in Figure 2, the TCEM is applied to bi-temporal feature pairs at four different scales to capture multi-scale change object information and provide preliminary retrieval guidance for subsequent detection tasks.

E. U-fusion Change Perception Module (UCPM)

After obtaining change features at different scales, how to aggregate their contextual information and guide it into the up-sampling process is the key to perceiving multi-scale changing geographical features. For the scale features extracted by the encoder at different levels, shallow features contain information such as color and texture, which helps in perceiving change boundaries. Deep features contain rich semantic information, which aids in identifying changing objects. Therefore, the U-fusio change perception module (UCPM) is used to bridge the semantic gap between multi-scale features, aggregating the global contextual information of the image. The UCPM consists of a multi-scale down-sampling feature aggregation block (MDFA) and a multi-scale up-sampling feature aggregation block (MUFA). By employing a U-shaped architecture to bidirectionally aggregate

contextual information across different levels, the ability to perceive changing objects at various scales is enhanced.

Specifically, in the multi-scale down-sampling block, for the features $\mathbf{F}_{c,i} \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times C_i'}$ and $\mathbf{F}_{c,i+1} \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 2C_i'}$ after the TCEM from two adjacent levels, the feature map sizes are first unified through feature reshaping. As shown in Fig. 2, for $\mathbf{F}_{c,i}$, a reorganized feature map $\hat{\mathbf{F}}_{c,i} \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 4C_i'}$ is obtained by sampling every other pixel along the horizontal and vertical directions on a per-channel basis. The information contained in every four adjacent channels in $\hat{\mathbf{F}}_{c,i}$ corresponds to the information of a single channel in $\mathbf{F}_{c,i}$. Therefore, each four adjacent channels in $\hat{\mathbf{F}}_{c,i}$ are convoluted in the form of group convolution, and the original feature representation is aggregated.

$$\mathbf{F'}_{c,i} = \sigma\left(\mathrm{BN}\left(\mathrm{GroupConv}_{3\times3}\left(\hat{\mathbf{F}}_{c,i}\right)\right)\right) \tag{13}$$

where $\mathbf{F'}_{c,i} \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 2C_i'}$, the number of groups in the group convolution is C_i'.

The feature reshaping operation to some extent avoids the information loss caused by down-sampling operations such as pooling layers and vanilla convolution. Then, after adding the features $\mathbf{F'}_{c,i}$ and $\mathbf{F}_{c,i+1}$ with the same dimension, the context information is preliminarily aggregated through a residual convolution block as:

$$\mathbf{F}_a = \mathbf{F'}_{c,i} \oplus \mathbf{F}_{c,i+1} \tag{14}$$

$$\mathbf{F'}_a = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times3}\left(\mathbf{F}_a\right)\right)\right) \oplus \mathbf{F}_a \tag{15}$$

where $\mathbf{F'}_a \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 2C_i'}$ represents the preliminarily aggregated feature.

To efficiently capture the dependencies between channels while minimizing the loss in dimensionality, an efficient channel attention (ECA) layer [53] is used to better learn the change semantics contained between channels, thereby optimizing the feature representation of $\mathbf{F'}_a$. ECA employs a non-diminishing, local cross-channel interaction strategy that significantly reduces the complexity of conventional channel attention computations while maintaining the ability to capture important channels.

Finally, the output from the ECA layer serves as the input for the next MDFA block, continuing the feature aggregation and down-sampling operations with $\mathbf{F}_{c,i+2}$. In the upsampling section, the structure of the MUFA block is similar to that of the MDFA block, with the only difference being that the feature reorganization down-sampling operation is replaced with feature reorganization up-sampling. Ultimately, the optimized change features from each layer are fed into a simple decoder and classification head via skip connections to obtain the final change detection prediction results $\mathbf{P} \in \mathbb{R}^{H\times W\times 2}$.

F. Loss Function

For the binary change detection tasks, it can actually be considered as a binary classification task. Therefore, the cross-entropy loss function is often used as the loss function for model training, which can be formulated as:

$$\mathcal{L}_{ce}\left(\mathbf{Y}, \mathbf{P}\right) = -\frac{1}{H_0 \times W_0} \sum_{i,j} \mathbf{Y}_{i,j} \log \mathbf{P}_{i,j} \tag{16}$$

where $\mathbf{Y}$ is the ground truth change detection map of the input bi-temporal images. H_0 and W_0 are the height and width of the bi-temporal images, respectively.

IV. Experiments

In this section, three public change detection datasets utilized for the experiments are discussed first. Then the relevant experimental details are introduce. Finally, experimental results comparing FIMP with the state-of-the-art methods and the ablation study results of FIMP are presented.

A. Dataset Introduction

The Sun Yat-Sen University CD Dataset (SYSU-CD) [54], High-resolution Complex Urban Scene CD Dataset (HRCUS-CD) [55], and Wuhan University Building CD Datasets (WHU-CD) [56] were selected for experimental validation. These datasets encompass various types of changes, such as building alterations, vegetation changes, and road construction, allowing for a comprehensive evaluation of the algorithm's performance.

1) SYSU-CD: This dataset comprises a total of 800 pairs of aerial images with a resolution of 0.5 meters. It documents the various changes that occurred in Hong Kong, China, from 2014 to 2017, including maritime construction, vegetation changes, road reconstruction, etc. The authors crop the original data into non-overlapping image patches sized 256 × 256 and divide them into training, validation, and test sets in a 6:2:2 ratio. Ultimately, the training, validation, and test sets include 12 000, 4 000, and 4 000 pairs of images, respectively.

2) HRCUS-CD: This dataset contains a total of 11,388 pairs of high-resolution remote sensing images with a resolution of 0.5 meters, each sized 256 × 256. It documents the changes that occurred in Zhuhai City, China, from 2010 to 2018 and from 2019 to 2022. The dataset primarily focuses on various types of building changes in complex scenes. The authors divide the original data into training, validation, and test sets in a 7:2:1 ratio. Ultimately, the training, validation, and test sets include 7 974, 2 276, and 1 138 pairs of images, respectively.

3) WHU-CD: This dataset comprises a single pair of aerial images with a resolution of 0.2 meters, each sized 32 207 × 15 354. It documents the changes in Christchurch, New Zealand, from the aftermath of the earthquake in 2011 to the reconstruction in 2012. This dataset primarily focuses on changes in buildings. For the convenience of model training, the original data was cropped into non-overlapping image patches sized 256 × 256 and randomly divided. Ultimately, the training, validation, and test sets include 5 950, 742, and 742 pairs of images, respectively.

B. Experimental Details

All experiments using the Pytorch platform, and the model is trained and tested on a workstation using a single NVIDIA GeForce RTX 3090 GPU with 24G of memory. Data augmentation on the training set is performed by randomly flipping and exchanging the order of the bi-temporal images. AdamW is used as the optimizer to optimize the model parameters, with a weight decay of 0.01 and an initial learning rate of

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

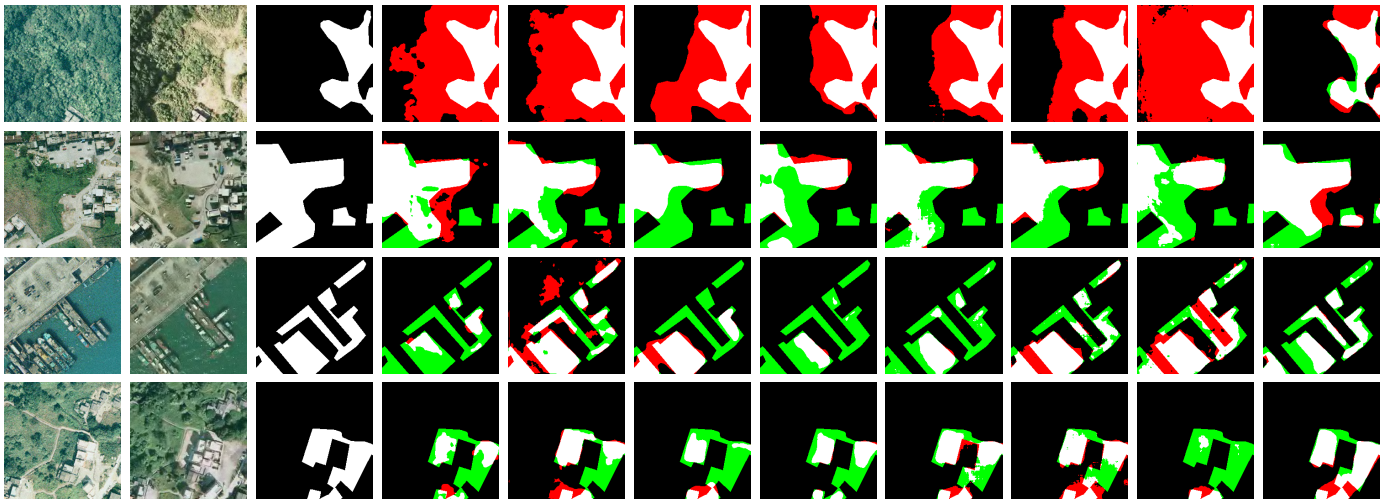


Fig. 4. Qualitative results of the proposed method compared with other methods on the SYSU-CD dataset. (a)T1 image; (b) T2 image; (c) Ground truth; (d) BIT [18]; (e) DSAMNet [54]; (f) TFI [57]; (g) AFCF3D-Net [58]; (h) P2V-CD [59]; (i) EATDer [60]; (j) MAFGNet [61]; and (k) FIMP. In the prediction results, white pixels represent true positives; black pixels represent true negatives; red pixels represent false positives; and green pixels represent false negatives.

0.001. A cosine annealing strategy is employed to adjust the learning rate, with the minimum learning rate set to 0.0001. The training is set for 200 epochs, with a batch size of 32.

Five commonly used metrics in change detection tasks, including the overall accuracy (OA) , precision (Pre), recall (Rec) , F1 score (F1), and Intersection over Union (IoU), are employed for quantitative analysis [42].

C. Comparison With State-of-the-Art Methods

To verify the effectiveness of FIMP, it is compared with seven methods: BIT [18], DSAMNet [54], TFI [57], AFCF3D-Net [58], P2V-CD [59], EATDer [60], and MAFGNet [61]. All comparative methods were implemented using their respective open-source codes and were tested according to the parameter settings provided in the original papers. Among them, BIT [18], DSAMNet [54], TFI [57], and AFCF3DNet [58] all use ResNet18 [62] as their backbone. P2V-CD [59], EATDer [60], and MAFGNet [61] utilize backbones that are modules specifically proposed in their respective methods. These methods are all representative or state-of-the-art in the field of change detection in recent years. Comparing them effectively shows the superior performance of FIMP.

Among these methods, BIT [18] represents bi-temporal images as semantic tokens. It uses a transformer encoder to model the contextual information of the images and refines the feature representation through a transformer decoder to achieve change detection. DSAMNet [54] enhances the encoder's ability to capture change information through deep supervision and integrates convolutional attention blocks into the metric module to refine the change detection results. TFI [57] achieves good detection results in complex terrain distribution scenarios through the interaction between temporal features and a multi-level feature refinement module. AFCF3D-Net [58] utilizes 3D convolution for feature extraction and fusion, thereby reducing the semantic gap between adjacent multi-scale features. P2V-CD [59] extends bi-temporal images into video frame sequences to address the issue of insufficient temporal modeling by the model. It fully exploits the geometric information and temporal correlations of images through spatial and temporal convolutional modules. EATDer [60] utilizes change boundary information to guide the transformer decoder in capturing long-distance contextual information, accurately identifying change objects while precisely locating their boundaries. MAFGNet [61] captures global contextual information of images using graph convolutional neural networks. It integrates multi-scale feature information by fusing features extracted from spatial graph convolutional networks and channel graph convolutional networks.

1) Qualitative Evaluation: The qualitative results on the three datasets are shown in Figs. 4-6. In the prediction maps of various methods, white pixels represent true positives; black pixels represent true negatives; red pixels represent false positives; and green pixels represent false negatives.

From Fig. 4, it can be observed that due to the uncertainty of imaging conditions, the color, brightness, and other style representations of bi-temporal images in SYSU-CD are significantly inconsistent. Among the seven comparison methods, AFCF3D-Net and P2V-CD achieved relatively better detection results. This may be attributed to the integration of spatial and temporal features in these two methods, which enhances the network's ability to capture change regions. However, both methods still exhibit a higher number of false positives in pseudo-change areas. As seen from the first row, FIMP effectively mitigates the impact of pseudo-changes caused by inconsistencies in vegetation phenotypes, especially in cases where other methods produce a large number of false positives. This might be due to the strategy of frequency interaction, which captures more precise change semantics while mitigating style differences. From the visualization results of the second to fourth rows, it can be observed that the proposed method is better at extracting morphological information of changing objects and is also capable of identifying smaller detection targets when facing multi-scale changes. This might be because FIMP, through its U-fusion approach, bridges the

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

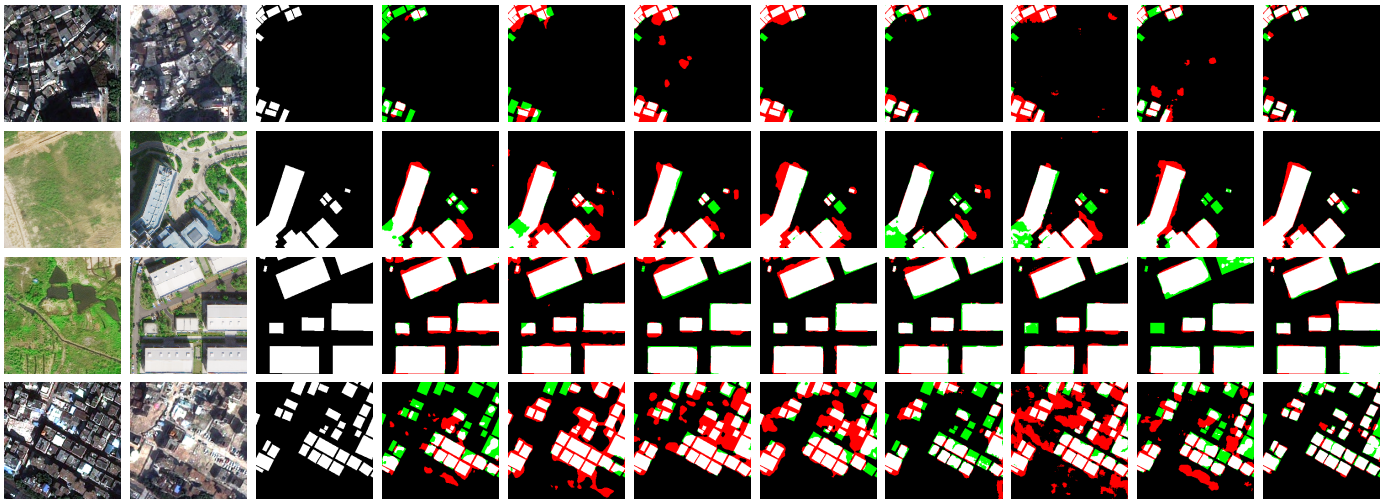


Fig. 5. Qualitative results of the proposed method compared with other methods on the HRCUS-CD dataset. (a) T1 image; (b) T2 image; (c) Ground truth; (d) BIT [18]; (e) DSAMNet [54]; (f) TFI [57]; (g) AFCF3D-Net [58]; (h) P2V-CD [59]; (i) EATDer [60]; (j)MAFGNet [61]; and (k) FIMP. In the prediction results, white pixels represent true positives; black pixels represent true negatives; red pixels represent false positives; and green pixels represent false negatives.

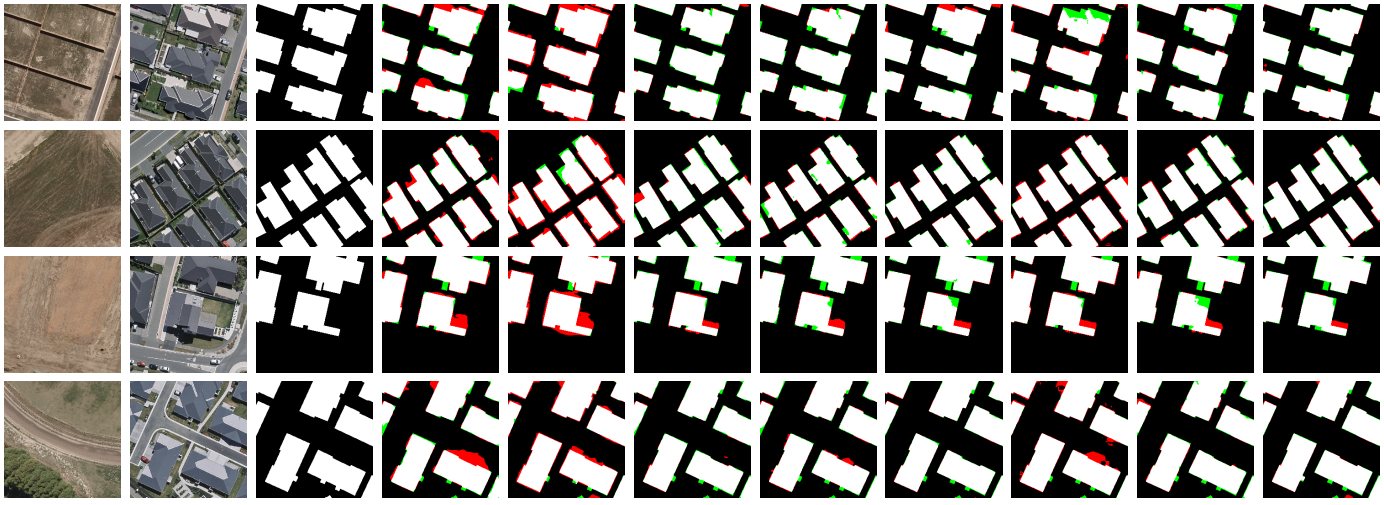


Fig. 6. Qualitative results of the proposed method compared with other methods on the WHU-CD dataset. (a) T1 image; (b) T2 image; (c) Ground truth; (d) BIT [18]; (e) DSAMNet [54]; (f) TFI [57]; (g) AFCF3D-Net [58]; (h) P2V-CD [59]; (i) EATDer [60]; (j)MAFGNet [61]; and (k) FIMP. In the prediction results, white pixels represent true positives; black pixels represent true negatives; red pixels represent false positives; and green pixels represent false negatives.

semantic gap between feature maps of different scales while fully aggregating global context information, thereby reducing the missed and false detections of change areas.

From Fig. 5, it can be observed that, compared to SYSU-CD, the changing objects in HRCUS-CD exhibit dense and small characteristics, and the change scenarios are more complex, making accurate identification of terrain changes even more challenging.

When addressing small and densely clustered change regions, P2V-CD demonstrates fewer false detections compared to other methods but suffers from a higher rate of missed detections. This is likely due to its robust modeling of temporal correlations, allowing for more accurate capture of change semantics, though it lacks sensitivity to multi-scale objects. TFI, AFCF3D-Net, and EATDer detect more complete change regions, but they also produce excessive false positives, likely because they aggregate multi-scale change features to some extent. Compared to these methods, FIMP achieves a better balance between false positives and missed detections, resulting in superior detection performance overall. Furthermore, from the results of the second and third rows, it can be seen that when facing pseudo change effects such as building shadows caused by the uncertainty of imaging conditions, FIMP is able to more meticulously delineate the true change boundaries in pseudo change areas compared to other methods. This may be due to the Fourier feature interaction strategy effectively filters out some pseudo-changes, and TCEM fully simulates the temporal correlation of the bi-temporal image.

WHU-CD is a dataset focused solely on building changes. As can be seen from Fig. 6, the morphological features of buildings are diverse, which poses a challenge to accurately depicting the boundaries of changed buildings. Compared to other methods, FIMP can more accurately define change boundaries, reducing false detections while more completely

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT CHANGE DETECTION METHODS ON SYSU-CD, HRCUS-CD, AND WHU-CD DATASETS, RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN RED AND THE SECOND BEST RESULTS ARE BLUE. ALL RESULTS OF THE THREE EVALUATION METRICS ARE DESCRIBED AS PERCENTAGES (%).

Method	SYSU-CD					HRCUS-CD					WHU-CD				
	OA	F1 Score	IoU	Pre	Rec	OA	F1 Score	IoU	Pre	Rec	OA	F1 Score	IoU	Pre	Rec
BIT [18]	89.63	77.93	63.84	78.25	77.61	98.86	67.54	50.99	72.31	63.35	98.85	87.31	77.47	85.36	89.35
DSAMNet [54]	87.60	75.59	60.76	70.57	81.39	98.74	66.22	49.50	66.66	65.78	98.48	84.40	73.01	77.56	92.56
TFI [57]	92.41	83.68	71.94	84.83	82.56	99.06	75.00	60.01	74.58	75.43	99.42	93.44	87.68	93.44	93.01
AFCF3D-Net [58]	92.04	82.95	70.87	83.76	82.16	98.99	74.01	58.74	72.06	76.07	99.42	93.40	87.62	94.58	92.25
P2V-CD [59]	90.36	78.86	65.10	81.66	76.25	98.99	71.16	55.23	76.85	66.25	99.44	93.56	87.90	95.18	92.00
EATDer [60]	91.03	80.96	68.01	81.02	80.90	98.62	67.25	50.66	60.56	75.61	99.03	89.27	80.62	87.85	90.74
MAFGNet [61]	91.07	80.69	67.62	82.36	79.08	98.80	63.97	47.02	73.19	56.81	99.29	91.99	85.17	91.87	92.12
FIMP	92.69	84.00	72.41	86.78	81.38	99.19	77.67	63.49	80.49	75.04	99.58	95.29	90.99	96.88	93.74

recognizing change areas. FIMP achieves better detection results by deeply interacting between bi-temporal images and fully aggregating multi-scale features, thereby narrowing the semantic gap while aggregating contextual information.

2) Quantitative Evaluation: The quantitative results on the three experimental datasets are shown in TABLE I. In these tables, red indicates the highest scores for each quantitative metric, while blue denotes the second highest scores.

It can be observed that FIMP achieves the best results on the three datasets. As an example of a more comprehensive evaluation metric, on SYSU-CD, HRCUS-CD, and WHU-CD, compared to the second highest F1 score, FIMP improved by 0.32%, 2.67%, and 1.73%, respectively. Compared to the second highest IoU, FIMP improved by 0.47%, 3.48%, and 3.09%, respectively. For BIT and MAFGNet, although they capture long-range semantic information in different ways, they neglect the effective utilization of change features at different levels. Therefore, their detection performance is generally mediocre when facing complex scenes. For DSAMNet, P2V-CD, and EATDer, although they can extract more representative change features, they merely perform a simple fusion of multi-level features, ignoring the impact of the semantic gap between deep and shallow features. Among all comparison methods, TFI and AFCD3D-Net also achieve good results on various datasets. This might be because they also fully aggregate multi-level features of different scales, capturing more contextual information while narrowing the semantic gap between deep and shallow features. However, these two methods do not consider the impact of style differences between bi-temporal images and directly interact with the extracted bi-temporal features. Compared to other methods, FIMP uses frequency interaction, employing adaptive frequency filtering to mine information relevant to the change detection task and adjust the global representation of bi-temporal features, thereby mitigating the impact of style differences between bi-temporal images to some extent.

D. Computational Complexity Analysis

To further evaluate the model size and computational complexity, the comparison results of the proposed method with the other seven comparison methods in terms of FLOPs and the number of parameters is shown in TABLE II. The best results for each metric are highlighted in bold in the TABLE II. w/o indicates that no pre-trained backbone is used, but is replaced by carefully designed modules in their paper. It can be observed that FIMP achieves better detection results with the smallest number of parameters and the least computational complexity, further validating the performance of the proposed method.

TABLE II
COMPARISON OF FLOPS AND PARAMETERS BETWEEN THE PROPOSED METHOD AND SEVEN COMPARISON METHODS.

Method	Backbone	Params (M)	FLOPs (G)
BIT [18]	Resnet18	3.04	10.90
DSAMNet [54]	Resnet18	16.95	72.31
TFI [57]	Resnet18	28.37	9.67
AFCF3D-Net [58]	Resnet18	17.64	31.71
P2V-CD [59]	w/o	5.42	33.03
EATDer [60]	w/o	6.59	23.46
MAFGNet [61]	w/o	6.58	62.18
FIMP	EfficientNet-b4	**2.45**	**2.85**

E. Ablation Studies and Parameter Analysis

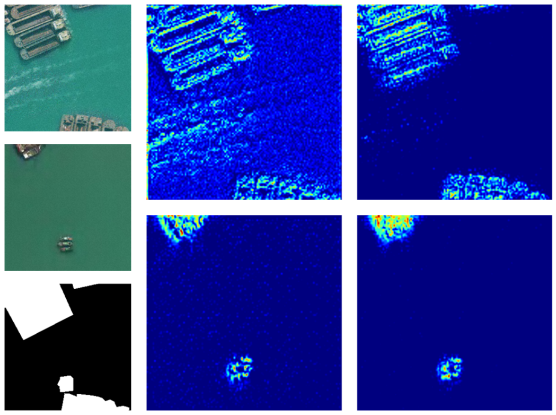
To verify the effectiveness of each module in FIMP, ablation studies were conducted on FIMP across three datasets under the same experimental conditions. The number of layers in UCPM is set to 3 by default. In the baseline model, the TCEM is replaced with a concatenation operation and a single 1×1 convolution layer that includes batch normalization and a PReLU activation function. The quantitative results of the ablation study in terms of F1 Score and IoU are summarized in TABLE III.

1) Effectiveness of Fourier feature interaction strategy: The Fourier feature interaction strategy optimizes the bi-temporal feature representation by transforming the original features into the frequency domain and adaptively weighting different frequency components. This approach effectively mitigates the impact of style differences while mining change information. From the first and second rows of TABLE III, it can be seen that when only the AFFM is added to the baseline model, the F1 scores on the three datasets increase by 1.24%, 0.33%, and 0.26%, respectively, and the IoU increase by 2.65%, 0.42%, and 0.46%, respectively. Furthermore, when only removing the AFFM from the complete FIMP, the F1 Score and IoU on the three datasets decrease by 0.83%, 1.23%, 0.39% and 1.21%, 1.62%, 0.70%, respectively.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

TABLE III
ABLATION STUDY RESULTS OF AFFM, TCEM, AND UCPM IN FIMP ON THREE DIFFERENT DATASETS.

AFFM	TCEM	UCPM	SYSU-CD		HRCUS-CD		WHU-CD	
			F1 Score	IoU	F1 Score	IoU	F1 Score	IoU
×	×	×	0.8054	0.6742	0.7504	0.6006	0.9440	0.8940
✓	×	×	0.8170	0.6907	0.7537	0.6048	0.9466	0.8986
×	✓	×	0.8249	0.7020	0.7570	0.6090	0.9476	0.9005
×	×	✓	0.8250	0.7022	0.7567	0.6087	0.9478	0.9008
✓	✓	×	0.8272	0.7054	0.7658	0.6206	0.9509	0.9065
✓	×	✓	0.8295	0.7088	0.7631	0.6169	0.9494	0.9037
×	✓	✓	0.8317	0.7120	0.7644	0.6187	0.9490	0.9029
✓	✓	✓	**0.8400**	**0.7241**	**0.7767**	**0.6349**	**0.9529**	**0.9099**



(a)

(b)

Fig. 7. Visualization results of feature maps on the SYSU-CD dataset (for the change category). Red and blue respectively represent areas that are highly relevant and lowly relevant to the change category. In (a) and (b), the first column from top to bottom represents the T1 image, the T2 image, and the ground truth; the second column represents the bi-temporal feature maps before frequency filtering; the third column represents the feature maps after bi-temporal frequency filtering.

To further demonstrate the effectiveness of the AFFM, class activation mapping (CAM) is utilized to visualize the bi-temporal feature maps before and after filtering. CAM intuitively demonstrates through heatmaps how the model perceives certain areas to be highly relevant to specific categories. The visualization results are displayed in Fig. 7. In the visualized feature maps, areas that the model considers to be more related to the change category appear redder, while areas with lower relevance to the change category appear bluer.

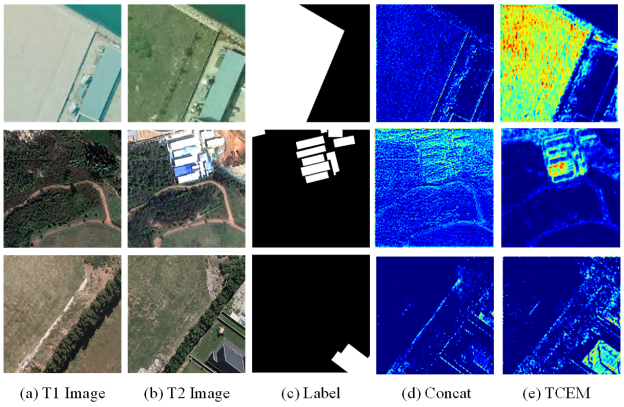


Fig. 8. Partial visualization results of feature maps from the ablation study of TCEM on three datasets. From top to bottom, the three rows are from the SYSU-CD, HRCUS-CD, and WHU-CD datasets, respectively. (d) represents the results of replacing TCEM with concatenation and 1×1 convolution.

From the first and second columns of Fig. 7 (a), it can be observed that before processing with AFFM, the wakes of ships on the sea or waves, as well as non-changing areas with inconsistent styles in bi-temporal images, are identified by the network as regions related to the change category. From the first and second columns of Fig. 7 (b), it can be seen that there is a significant color inconsistency in the sea surface part of the bi-temporal images, causing the encoder to mistakenly identify the sea surface as a change area, which will severely affect the training of the model. However, after passing through the AFFM module, the aforementioned pseudo changes are effectively filtered out by AFFM. At the same time, the model's attention to the actual change areas is increased. This further shows the effectiveness of mitigating the impact of pseudo changes such as style differences through AFFM.

2) Effectiveness of the temporal change enhancement module: TCEM fully interacts the bi-temporal features optimized after filtering to model temporal correlations and highlight change areas. From the first and third rows of TABLE III, it can be seen that when only TCEM is added to the baseline model, the F1 scores on the three datasets increase by 1.95%, 0.66%, and 0.36%, respectively, and the IoU scores increase by 2.78%, 0.84%, and 0.65%, respectively. From the last and the third-to-last rows of Table 5, it can be observed that when TCEM is solely removed from the complete FIMP, the F1 scores and IoU on the three datasets decrease by 1.05%,

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3500073

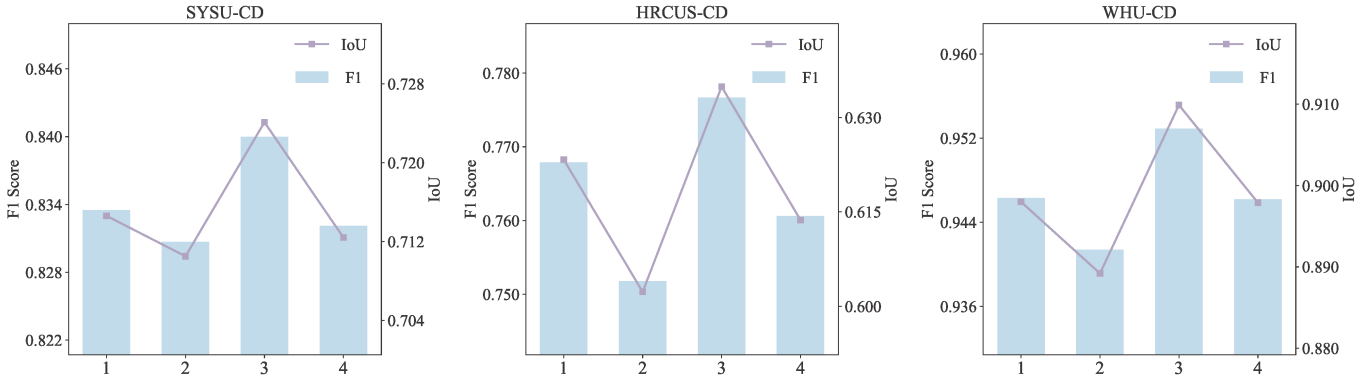


Fig. 9. Experimental results of UCPM with different numbers (1,2,3,4) of layers on the three datasets. The blue bars represent the F1 Score, and the purple line represents the IoU.

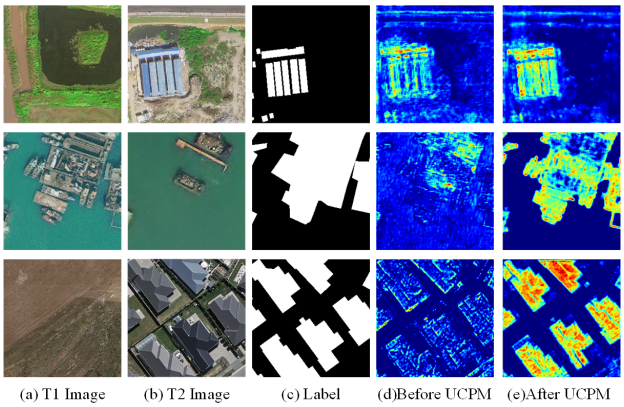


Fig. 10. Partial visualization results of feature maps from the ablation study of UCPM on three datasets. From top to bottom, the three rows are from the SYSU-CD, HRCUS-CD, and WHU-CD datasets, respectively.

TABLE IV
ABLATION EXPERIMENTS OF FEATURE MIXING METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN **BLOD**. (UNDER F1-SCORE)

Mixing Method	SYSU-CD	HRCUS-CD	WHU-CD
$\mathbf{A}-\mathbf{B}$	83.27	75.57	95.09
$\lvert\mathbf{A}-\mathbf{B}\rvert$	82.78	76.38	94.48
Concat	82.95	76.31	94.93
Add	83.21	77.12	94.94
TCEM	**84.00**	**77.67**	**95.29**

1.36%, 0.35% and 1.53%, 1.80%, 0.62%, respectively.

From Fig. 8, it can be seen that the concatenation operation has poor recognition effect on change areas across the three datasets. This is because simple concatenation merely fuses the bi-temporal features together, making it difficult to model temporal correlations. However, TCEM enhances the recognition of change areas by utilizing the difference between bi-temporal features, under the premise of fully modeling temporal correlations, resulting in improved recognition effects.

Additionally, various feature mixing methods have been explored in greater detail. As shown in the TABLE IV, $\mathbf{A}-\mathbf{B}$ represents the subtraction of bi-temporal features, $|\mathbf{A}-\mathbf{B}|$ indicates the absolute value of the subtracted features, Concat refers to the concatenation of bi-temporal features along the channel dimension, and Add signifies the addition of the bi-temporal features. The best results are highlighted in bold. Experimental results indicate that while these simple feature fusion methods offer advantages for different datasets, none of them fully capture the semantic relationships between the bi-temporal features. Among these methods, TCEM achieved the highest F1 score, indicating that it effectively models the temporal correlations between bi-temporal features.

3) Effectiveness of the U-fusion change perception module: UCPM enhances the model's ability to capture change objects of different scales by bidirectionally aggregating features of various levels through a U-shaped architecture. To verify the effectiveness of UPCM, when it is added to the baseline model individually, it achieved F1 Scores of 82.50%, 75.67%, and 94.78% and IoU of 70.22%, 60.87%, and 90.08% on the three datasets, respectively. Compared to the baseline network, the F1 Scores increased by 2.04%, 0.63%, and 0.38%, and the IoU increased by 2.82%, 0.81%, and 0.68%, respectively. When only UCPM is removed, the F1 scores and IoUs on the three datasets decrease by 1.28%, 1.09%, 0.20% and 1.87%, 1.43%, 0.34%, respectively.

From Fig. 10, it can be seen that UCPM effectively aggregates the contextual information contained in multi-scale features, enhancing the model's ability to perceive various types of change objects. After processing with UCPM, the model is not only able to recognize change objects of different scales more completely but also accurately delineate the boundaries of changes.

To investigate the effect of UCPM layer count on detection performance, experiments were conducted with 1, 2, 3, and 4 layers across three datasets, with F1 Score and IoU employed as quantitative evaluation metrics. As depicted in Fig. 9, the detection performance exhibits an increasing trend as the number of UCPM layers grows, reaching its peak with 3 layers. When fewer layers are utilized, the UCPM's capacity for multi-scale feature fusion is constrained, limiting its effectiveness in enhancing change detection. Conversely, an excessive number of layers leads to a decline in performance, likely due to overfitting. Therefore, 3 UCPM layers provide an optimal balance between performance and computational efficiency in the overall model.

TABLE V
ABLATION EXPERIMENTS OF DIFFERENT SCALES IN UCPM.THE BEST RESULTS ARE HIGHLIGHTED IN **BLOD**. (UNDER F1-SCORE)

Case	SYSU-CD	HRCUS-CD	WHU-CD
Case1	82.70	74.28	94.45
Case2	83.00	76.40	94.51
Case3	82.99	76.22	94.76
Case4	**84.00**	**77.67**	**95.29**

V. DISCUSSION

To further investigate the effectiveness of aggregating multi-scale features in UCPM, we designed experiments under the following four cases. Case1: UCPM is not utilized, and only the deepest layer features are progressively upsampled to generate the prediction results. Case2: Only the deepest layer MDFA and MUFA within UCPM are used. Case3: Only the second and third layer MDFA and MUFA within UCPM are employed. Case4: The complete UCPM is applied. As presented in Table V, the F1 score is the lowest without the use of UCPM (Case1). When UCPM is applied to aggregate partial scale features (Case2 and Case3), the F1 score shows a certain degree of improvement. This could be attributed to the fact that deep semantic features alone are insufficient to accurately capture change details across different scales. The F1 score for Case4 is the highest, further indicating that UCPM effectively bridges the semantic gap between multi-scale features through multi-layer bidirectional aggregation, thereby enhancing the network's ability to capture complex change objects.

In addition, we also conducted an analysis of the limitations of the proposed method. Fig. 11 illustrates three failure cases of FIMP on the SYSU-CD, HRCUS-CD, and WHU-CD datasets. For the example in the first row of Fig. 11, FIMP has identified most of the change areas, but there are still incomplete detections, with the missing regions highlighted in green. This may be due to the change from grassland to trees in the bi-temporal images, where the features of both are quite similar, making it difficult for the model to accurately distinguish them. For the HRCUS-CD and WHU-CD datasets, which primarily focus on building changes, the examples in the second and third rows show that FIMP can relatively accurately locate the change areas and outline the change boundaries. However, due to the similarities between buildings and the ground in some features, the model still exhibits some false positives and false negatives in certain areas. In our opinion, incorporating background feature learning to enhance the model's ability to handle complex backgrounds could be a viable solution. In future work, we will further explore this direction.

VI. CONCLUSION

In this paper, a change detection method for high-resolution remote sensing images based on Fourier feature interaction and multi-scale perception is proposed (FIMP). Initially, FIMP uses the Fourier feature interaction strategy to enhance bi-temporal feature representation. Through the adaptive frequency filtering module (AFFM), it effectively reduces the impact of pseudo changes like style discrepancies caused by varying imaging conditions. Subsequently, a temporal change enhancement module (TCEM) is used to model the temporal correlations between bi-temporal features, which captures change information while highlighting change areas. Moreover, for the rich semantic information contained in feature maps of different scales, a U-fusion change perception module (UCPM) is employed. It aggregates contextual information between multi-scale change features while narrowing the semantic gap between features at different levels. Conclusively, experiments on three publicly available change detection datasets show that FIMP surpasses seven most advanced change detection methods. Future research will further focus on weakly supervised change detection methods.

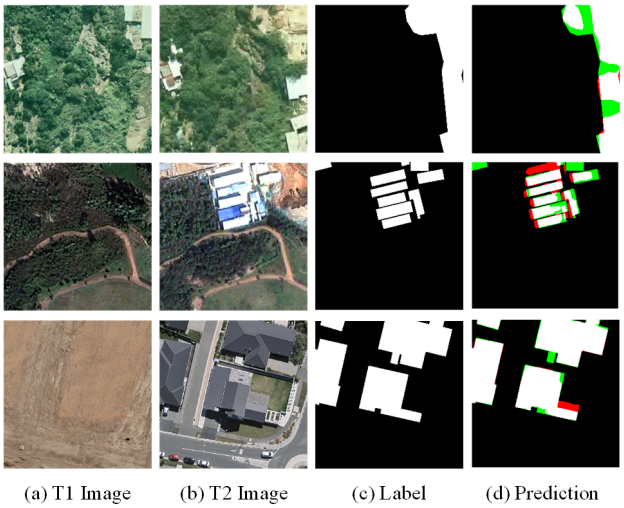


Fig. 11. Failure examples obtained by FIMP on three datasets. From top to bottom, the three rows are from the SYSU-CD, HRCUS-CD, and WHU-CD datasets, respectively. In the prediction results, white pixels represent true positives; black pixels represent true negatives; red pixels represent false positives; and green pixels represent false negatives.

REFERENCES

[1] J. Wang, Y. Zhong, and L. Zhang, "Change detection based on supervised contrastive learning for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[2] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, 2023.

[3] H. Fang, S. Guo, X. Wang, S. Liu, C. Lin, and P. Du, "Automatic urban scene-level binary change detection based on a novel sample selection approach and advanced triplet neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.

[4] E. Hamidi, B. G. Peter, D. F. Muñoz, H. Moftakhari, and H. Moradkhani, "Fast flood extent monitoring with sar change detection using google earth engine," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–19, 2023.

[5] Z. Lv, J. Liu, W. Sun, T. Lei, J. A. Benediktsson, and X. Jia, "Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[6] L. Miao, X. Li, X. Zhou, L. Yao, Y. Deng, T. Hang, Y. Zhou, and H. Yang, "Snunet3+: A full-scale connected siamese network and a dataset for cultivated land change detection in high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2023.

[7] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, 2006.

[8] C. Yang, J. H. Everitt, and J. M. Bradford, "Yield estimation from hyperspectral imagery using spectral angle mapper (sam)," *Trans. ASABE*, vol. 51, no. 2, pp. 729–737, 2008.

[9] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, pp. 772–776, 2009.

[10] J. Wang and C.-I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, 2006.

[11] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, 2009.

[12] 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, pp. 1–15, 2023.

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

[14] S. Zhao, X. Zhang, P. Xiao, and G. He, "Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[15] Y. Liu, F. Zhang, S. Zhang, K. Zhang, J. Sun, and L. Bruzzone, "Content-guided spatial-spectral integration network for change detection in hr remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[16] H. Chen, H. Zhang, K. Chen, C. Zhou, S. Chen, Z. Zou, and Z. Shi, "Continuous cross-resolution remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–20, 2023.

[17] K. Jiang, J. Liu, W. Zhang, F. Liu, and L. Xiao, "Manet: An efficient multi-dimensional attention-aggregated network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.

[18] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[19] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "Saan: Similarity-aware attention flow network for change detection with vhr remote sensing images," *IEEE Trans. Image Process.*, vol. 33, pp. 2599–2613, 2024.

[20] Z. Li, S. Cao, J. Deng, F. Wu, R. Wang, J. Luo, and Z. Peng, "Stadecdnet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.

[21] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention siamese network for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[22] X. Xu, Z. Yang, and J. Li, "Amca: Attention-guided multi-scale context aggregation network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–19, 2023.

[23] Y. Zhang, Y. Zhao, Y. Dong, and B. Du, "Self-supervised pre-training via multi-modality images with transformer for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[24] B. Jiang, Z. Wang, X. Wang, Z. Zhang, L. Chen, X. Wang, and B. Luo, "Vct: Visual change transformer for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[25] Q. Xu, Y. Shi, J. Guo, C. Ouyang, and X. X. Zhu, "Ucdformer: Unsupervised change detection using a transformer-driven image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

[26] G. Liu, Y. Yuan, Y. Zhang, Y. Dong, and X. Li, "Style transformation-based spatial–spectral feature learning for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2020.

[27] 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, pp. 1–17, 2022.

[28] J. Liu, W. Xuan, Y. Gan, Y. Zhan, J. Liu, and B. Du, "An end-to-end supervised domain adaptation framework for cross-domain change detection," *Pattern Recognit.*, vol. 132, p. 108960, 2022.

[29] C. Zhao, Y. Tang, S. Feng, Y. Fan, W. Li, R. Tao, and L. Zhang, "High resolution remote sensing bitemporal image change detection based on feature interaction and multi-task learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[30] Y. Liu, R. Yu, J. Wang, X. Zhao, Y. Wang, Y. Tang, and Y. Yang, "Global spectral filter memory network for video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Switzerlabd:Springer, 2022, pp. 648–665.

[31] Z. Li, H. Lyu, and J. Wang, "Fusionu-net: U-net with enhanced skip connection for pathology image segmentation," in *Asian Conference on Machine Learning*. PMLR, 2024, pp. 694–706.

[32] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[33] Y. Wen, X. Ma, X. Zhang, and M.-O. Pun, "Gcd-ddpm: A generative change detection model based on difference-feature guided ddpm," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[34] L. Mei, Z. Ye, C. Xu, H. Wang, Y. Wang, C. Lei, W. Yang, and Y. Li, "Scd-sam: Adapting segment anything model for semantic change detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.

[35] S. Dong, L. Wang, B. Du, and X. Meng, "Changeclip: Remote sensing change detection with multimodal vision-language representation learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 208, pp. 53–69, 2024.

[36] P. Duhamel and M. Vetterli, "Fast fourier transforms: a tutorial review and a state of the art," *Signal processing*, vol. 19, no. 4, pp. 259–299, 1990.

[37] B. Qin, S. Feng, C. Zhao, B. Xi, W. Li, and R. Tao, "Fdgnet: Frequency disentanglement and data geometry for domain generalization in cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2024.

[38] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Rec.*, 2021, pp. 6891–6902.

[39] J. Guo, N. Wang, L. Qi, and Y. Shi, "Aloft: A lightweight mlp-like architecture with dynamic low-frequency transform for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Rec.*, 2023, pp. 24 132–24 141.

[40] L. Chen, L. Gu, D. Zheng, and Y. Fu, "Frequency-adaptive dilated convolution for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Rec.*, 2024, pp. 3414–3425.

[41] F. Zhang, A. Panahi, and G. Gao, "Fsanet: Frequency self-attention for semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 4757–4772, 2023.

[42] J. Liu, S. Li, R. Dian, Z. Song, and X. Kang, "Mdenet: Multi-domain differential excavating network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.

[43] H. Zheng, M. Gong, T. Liu, F. Jiang, T. Zhan, D. Lu, and M. Zhang, "Hfa-net: High frequency attention siamese network for building change detection in vhr remote sensing images," *Pattern Recognit.*, vol. 129, p. 108717, 2022.

[44] Y. Tang, S. Feng, C. Zhao, Y. Fan, Q. Shi, W. Li, and R. Tao, "An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2023.

[45] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[46] S. Lee, J. Bae, and H. Y. Kim, "Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Rec.*, 2023, pp. 11 776–11 785.

[47] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4085–4095.

[48] B. Cao, Q. Wang, P. Zhu, Q. Hu, D. Ren, W. Zuo, and X. Gao, "Multi-view knowledge ensemble with frequency consistency for cross-domain face translation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9728–9742, 2023.

[49] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 980–993, 2021.

[50] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 4479–4488, 2020.

[51] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[52] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[53] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proc. Conf. Comput. Vis. Pattern Rec.*, 2020, pp. 11 534–11 542.

[54] 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, pp. 1–16, 2022.
[55] J. Zhang, Z. Shao, Q. Ding, X. Huang, Y. Wang, X. Zhou, and D. Li, "Aernet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.
[56] 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, pp. 574–586, 2018.
[57] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
[58] 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, pp. 1–14, 2023.
[59] 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, 2022.
[60] J. Ma, J. Duan, X. Tang, X. Zhang, and L. Jiao, "Eatder: Edge-assisted adaptive transformer detector for remote sensing change detection," *IEEE Trans. Geosci. and Remote Sens.*, vol. 62, pp. 1–15, 2024.
[61] Y. Shangguan, J. Li, Z. Chen, L. Ren, and Z. Hua, "Multiscale attention fusion graph network for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024.
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

Yongqi Chen received the B.S. degree in communication engineering from the College of Information and Communication Engineering of Harbin Engineering University in 2022. He is currently pursuing the Ph.D. degree in information and communication engineering with Harbin Engineering University, Harbin, China.

The main research direction is remote sensing image change detection.

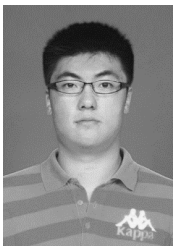
Shou Feng (Member, IEEE) is an associate Professor of Harbin Engineering University, China. He received his Ph.D. degree in 2019 from Harbin Institute of Technology, China.

His main research interests include remote sensing image processing, data mining, machine learning and etc.

Chunhui Zhao received the BS and MS degree from Harbin Engineering University, in 1986 and 1989, respectively, and his PhD degree in Department of Automatic Measure and Control at Harbin Institute of Technology in 1998.

He was a postdoctoral research fellow in the College of Underwater Acoustical Engineering of Harbin Engineering University. At present, he is working in the College of Information and Communication Engineering of Harbin Engineering University as a professor and doctoral supervisor.

Prof. Zhao is a senior member of Chinese Electronics Academy. His research interests include digital signal and image processing, mathematical morphology and hyperspectral remote sensing image processing.



Nan Su received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2017.

He is currently an Associate Professor with Harbin Engineering University, Harbin. His main research areas are remote sensing image processing and data mining.

Wei Li (Senior Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

Subsequently, he spent one year as a Postdoctoral Researcher with the University of California at Davis, Davis, CA, USA. He is a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. His research interests include hyperspectral image analysis, pattern recognition, and data reconstruction.

Dr. Li is currently an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Ran Tao (Senior Member, IEEE) was born in 1964. He received the B.S. degree from the Electronic Engineering Institute of PLA, Hefei, China, in 1985, and the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1990 and 1993, respectively.

In 2001, he was a Senior Visiting Scholar with the University of Michigan at Ann Arbor, Ann Arbor, MI, USA. He is a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He has three books and over 100 peer-reviewed journal articles. His research interests include fractional Fourier transform and its applications, theory, and technology for radar and communication systems.

Dr. Tao was a Distinguished Professor of the Changjiang Scholars Program in 2009. He is a member of the Wireless Communication and Signal Processing Commission of International Union of Radio Science (URSI). He is the Vice Chair of the IEEE China Council and the URSI China Council. He is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.

Jinchang Ren received the B.Eng., M.Eng., and D.Eng. degrees from Northwestern Polytechnical University, Xi'an, China, in 1992, 1997, and 2000, respectively, and the Ph.D. degree from the University of Bradford, Bradford, U.K., in 2019.

He is currently a Professor of computing with Robert Gordon University, Aberdeen, U.K. He has authored or coauthored more than 300 peer-reviewed journal articles or conference papers. His research interests include hyperspectral imaging, image processing, computer vision, big data analytics, and machine learning.