remote sensing

*Article*

# Bitemporal Remote Sensing Image Change Detection Network Based on Siamese-Attention Feedback Architecture

**Hongyang Yin [1], Chong Ma [1], Liguo Weng [1,*], Min Xia [1] and Haifeng Lin [2]**

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, B-DAT, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20211249112@nuist.edu.cn (H.Y.); 20211249062@nuist.edu.cn (C.M.); xiamin@nuist.edu.cn (M.X.)
[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210000, China; haifeng.lin@njfu.edu.cn
* Correspondence: 002311@nuist.edu.cn

**Abstract:** Recently, deep learning-based change detection methods for bitemporal remote sensing images have achieved promising results based on fully convolutional neural networks. However, due to the inherent characteristics of convolutional neural networks, if the previous block fails to correctly segment the entire target, erroneous predictions might accumulate in the subsequent blocks, leading to incomplete change detection results in terms of structure. To address this issue, we propose a bitemporal remote sensing image change detection network based on a Siamese-attention feedback architecture, referred to as SAFNet. First, we propose a global semantic module (GSM) on the encoder network, aiming to generate a low-resolution semantic change map to capture the changed objects. Second, we introduce a temporal interaction module (TIM), which is built through each encoding and decoding block, using the feature feedback between two temporal blocks to enhance the network's perception ability of the entire changed target. Finally, we propose two auxiliary modules—the change feature extraction module (CFEM) and the feature refinement module (FRM)—which are further used to learn the fine boundaries of the changed target. The deep model we propose produced satisfying results in dual-temporal remote sensing image change detection. Extensive experiments on two remote sensing image change detection datasets demonstrate that the SAFNet algorithm exhibits state-of-the-art performance.

**Keywords:** deep learning; remote sensing images; change detection

check for updates

**Citation:** Yin, H.; Ma, C.; Weng, L.; Xia, M.; Lin, H. Bitemporal Remote Sensing Image Change Detection Network Based on Siamese-Attention Feedback Architecture. *Remote Sens.* **2023**, *15*, 4186. https://doi.org/10.3390/rs15174186

Academic Editors: Xinzheng Zhang, Hong Huang and Ce Zhang

Received: 9 July 2023
Revised: 3 August 2023
Accepted: 7 August 2023
Published: 25 August 2023

## 1. Introduction

### 1.1. Background

With the development of geographic information technology, significant progress has been made in Earth observation technology, and various sensors and equipment have been widely applied. Remote sensing technology captures target objects through satellites equipped with sensors, analyzing the characteristic information of the target, extracting key features, and applying them [1–4]. At the same time, remote sensing technology can also capture spectral information beyond ultraviolet, infrared, and visible light to enrich spatial feature information. Due to these advantages, remote sensing images have gradually become a reliable and indispensable data source for obtaining surface information [5–9]. The change detection of dual-temporal remote sensing images refers to shooting the ground via remote sensing platforms such as satellites and drones, capturing images at different periods to locate areas that have changed and those that have not [10–12]. Simply put, it recognizes the differences in the status of geographical objects under different time distributions. Remote sensing image change detection is widely applied in geological surveying [13–15], environmental monitoring [16–18], resource management [19,20], urban expansion [21,22], and many other fields [23–26].

How to effectively extract true change information from dual-temporal remote sensing images is a crucial research direction at present. Over the past decades, many change detection methods have been proposed and applied. Images obtained in actual environments are subject to many uncertain factors, such as lighting, seasons, and shooting angles. Change detection methods can be divided into pixel-level change detection, feature-level change detection, and object-level change detection according to the difference in change detection elements. Pixel-level change detection considers each pixel as a basic unit and determines whether the values in the pixel have changed through comparison. Common methods include interpolation [27], the ratio method [28], and principal component analysis (PCA) [29]. However, these methods all have high requirements for the accuracy of image registration and are now mostly used as part of a framework, rather than used alone. Deng et al. [30] used PCA to enhance data and input the processed features into the classifier to detect land changes. Celik [31] proposed a detection method that combines PCA and K-means. Feature-level change detection extracts significant features from the original dual-temporal images and analyzes them for change detection. This method often extracts texture features and edge features. Zhang et al. [32] proposed a texture feature extraction method that can better describe texture features and spatial feature distribution based on local detail features. Guiming and Jidong [33] proposed an edge detection method that improves the Canny operator and achieves better noise-smoothing effects while retaining more details. Object-level change detection treats the image as a combination of objects with different semantic information and carries out semantic detection on these objects. Peng et al. [34] proposed a UNet++ based encoder–decoder structure that uses a multi-side output fusion method to obtain a higher-accuracy change map.

According to different technical means used, change detection methods can also be divided into traditional methods and methods based on deep learning. Traditional change detection generally consists of three steps: image preprocessing, difference map generation, and difference map analysis. In the image preprocessing stage, methods such as geometric correction and image registration are used to make the dual-temporal remote sensing images comparable in space and spectrum. In the difference map generation stage, a feature matrix representing the distance between dual-temporal remote sensing images is found through methods such as interpolation and ratio. In the difference map analysis stage, methods such as thresholding and clustering are combined to classify the pixels in the difference map. He et al. [35] proposed a dynamic threshold algorithm based on the merged fuzzy C-means algorithm to generate each pixel's membership value and a global initial threshold, which can reduce speckle noise and better retain detailed information. The main drawback of these methods is that they do not consider the surrounding pixel information when determining the correctness of the detected pixels, and only take the degree of fit between the statistical model and the actual data distribution as the standard. Thonfeld et al. [36] proposed robust change vector analysis (CVA), which considers the neighborhood information of each pixel to alleviate the effect of poor co-registration between images but does not capture object-level information. Zheng et al. [37] proposed a method based on combined difference images and K-means clustering. The use of mean and median filters ensures that edge information is well preserved while considering local consistency. Luppino et al. [38] proposed the use of a cycle-consistent generative adversarial network to transcode images from different sensors into the same domain in an unsupervised manner, and further implemented change detection through CVA. Traditional detection methods are complicated to operate and yield relatively low detection accuracy.

In recent years, the development of deep learning algorithms has provided new solutions for change detection, and it has also been widely applied in the research fields of object detection and image segmentation [39]. It builds a neural network structure to obtain a model framework, inputs the dataset into the network and then outputs the desired results according to the needs of the task, and the obtained model structure will have stronger robustness. Compared with traditional methods, deep learning-based methods can achieve higher scores, and also reduce some cumbersome steps of data preprocessing, and the end-to-end learning model framework is easier to be directly used. Gong et al. [40] first applied CNN to solve the task of remote sensing image change detection and achieved good results, proving the feasibility of CNN in the task of remote sensing image change detection. Subsequently, many researchers proposed CNN-based change detection methods. Zhan et al. [41] proposed a deep Siamese convolutional neural network to solve the problem of change detection, which extracted the change information of dual-temporal remote sensing images through the weight-sharing Siamese neural network, improving the operational efficiency of the model. On this basis, weight-sharing Siamese neural networks have been widely applied in the task of remote sensing image change detection. Zhang et al. [42] proposed a deep Siamese semantic network change detection method, improved the loss function, and used the triplet of piecewise functions to strengthen the robustness of the model. Chen and Shi [43] proposed a spatio-temporal attention neural network based on a connected body, dividing the image into sub-regions of multiple scales and introducing the self-attention mechanism in them, thus capturing spatio-temporal dependencies of various scales. Song et al. [44] proposed a change detection network based on the U-shaped structure, which extracts and learns the similar feature information, different feature information, and global feature information of dual-temporal remote sensing images through multiple branches. Wang et al. [45] proposed a hyperspectral image change detection method. The method first encodes the position of each pixel in the image, and then uses a spectral transform coder and a spatial transform coder to extract spectral sequence information and spatial texture information, respectively. Finally, the time-domain transformer is used to extract the useful change features between the current image pair, and the detection result is obtained through the multi-layer perceptron. Zhang et al. [46] proposed a cascaded attention-induced differential representation learning method for multispectral change detection, which explores the correlation of features extracted from bitemporal images to obtain more discriminative features, and finally detects the discriminative features and obtains the final detection map.

With the emergence of high-resolution remote sensing images, traditional remote sensing image change detection methods have been unable to solve related problems well. The ground object scenes in high-resolution remote sensing images are generally more complex, and the seasons or lighting conditions of different time-phase remote sensing images are different, which may lead to different spectral characteristics of ground objects with the same semantic concept at different times and different spatial positions, which introduces a lot of noise interference to remote sensing images, making it more difficult to detect changes in specific ground objects. The remote sensing image change detection method based on deep learning can better model the relationship between remote sensing images and real objects by virtue of its powerful representation ability. Wang et al. [47] used Faster R-CNN to detect changes in high-resolution remote sensing images. Experiments show that the detection accuracy of this method has been improved, and the detection probability of wrong samples has been reduced, thereby extracting more real changes information. Ding et al. [48] proposed a novel dual-branch end-to-end network to build change detection, and innovatively introduced a spatial attention mechanism-guided cross-layer addition and skip connection module to aggregate multi-level contextual information, weakening original image features and differential features heterogeneity among them, and direct the network's attention to regions where changes occur. Shu et al. [49] proposed a dual-perspective change context network for change detection, the process of extracting and optimizing change features by bitemporal feature fusion and context modeling. Yin et al. [50] proposed an attention-guided change detection Siamese network, which combines shallow spatial information and deep semantic information, to assist in the restoration of edge details of change areas and the reconstruction of small targets during upsampling.

### *1.2. Related Work*

Considering the collection problem of dual-temporal remote sensing images, potential factors such as sensors, illumination, and solar angle inevitably cause the same object to have different positions and spectra, causing visual changes in remote sensing images. However, these changes are not caused by actual changes in the ground objects and should not be detected. This redundant information in model training will bring up such a problem: if the previous block fails to correctly segment the entire target, the incorrect predictions may accumulate in subsequent blocks, leading to an inability to obtain structurally complete change detection results. Therefore, how to fully utilize the information contained in the dual-temporal remote sensing images and eliminate redundant useless information is a difficulty we are currently facing. In this paper, we propose a new change detection network in response to the above issues, referred to as a dual-temporal remote sensing image change detection network based on the Siamese-attention feedback system architecture (SAFNet). First, we propose a global semantic module (GSM) on the encoder network to generate a low-resolution semantic change map to roughly capture the changing objects and provide semantic guidance for the reconstruction of the change map. Then, we propose a temporal interaction module (TIM). Unlike previous methods that only use a single cascading operation or subtraction operation for dual-time feature fusion, TIM can enhance the interaction between dual-time features, filter out redundant information, and improve the network's perception of the entire changing target. Finally, we also propose a change feature extraction module (CFEM) to capture temporal difference information at different feature levels, and a feature refinement module (FRM) to adaptively focus on the change area, enhancing the network's detection ability of edge information and small targets. Our work's main contributions are summarized as follows:

1. We propose a bitemporal remote sensing image change detection network based on the Siamese-attention feedback system architecture (SAFNet) to address the challenges in change detection tasks. We design a temporal interaction module (TIM). When multi-scale features in the encoder block are passed into the corresponding decoder, the network's perception of the changing target is enhanced by using TIM to implement feature feedback between the two time steps, thus producing better detection results. SAFNet produces prediction outputs step by step and eventually obtains the best change prediction map.
2. We propose the global semantic module (GSM), change feature extraction module (CFEM), and feature refinement module (FRM). By introducing GSM into the deep layer of the encoder network to obtain context-aware semantic change information of multi-scale and multi-receptive fields, it can guide the network to better locate significant change areas during the learning process, and reduce network false detection and missed detection. CFEM extracts difference information from the features of dual-temporal remote sensing images between each level, better learning the edge features and texture features of the change features. FRM enables the network to capture change features in both spatial and channel dimensions, eliminates and suppresses redundant features, and weights the feature extraction for the next time step, thereby improving the network's detection accuracy.
3. Extensive experiments on two remote sensing image change detection datasets show that compared with other deep learning-based change detection algorithms, our proposed SAFNet network demonstrates robustness and high precision.

The rest of the paper is as follows: Section 2 introduces each module in the model and analyzes the composition of the dataset. Section 3 analyzes the performance of the model through experiments. Section 4 discusses the strengths and future of our approach. Section 5 summarizes our work in this paper.

## 2. Materials and Methods

In this section, we first give a detailed description of our proposed method, then introduce the datasets used in our experiments, and finally give the experimental details.

### *2.1. Proposed Approach*

In this paper, we propose a bitemporal remote sensing image change detection network based on Siamese-attention feedback system architecture (SAFNet) for predicting a change map with complete structure and fine boundaries. The following starts with the backbone network, and then details the specific implementation of each module.

### 2.1.1. Network Architecture

Since the purpose of the change detection task is to distinguish between the changed area and the non-changed area in the dual-temporal remote sensing image, in view of the weight-sharing twin structure of the Siamese network, the two images can be extracted separately and then differentiated. The idea is very suitable for the change detection task. We know that the attention mechanism can extract changing feature information from the spatial dimension and the channel dimension very well, and can effectively improve the efficiency of feature extraction. Based on this, we take ResNet34 [51,52] as our backbone network and build it with a decoder and encoder approach. Figure 1 shows the model's whole organizational structure. The decoder consists of five pairs of twin networks with shared weights, represented by $E_i^{(l)}$ ($l \in \{1,2,3,4,5\}$ represents each decoder block; $i \in \{1,2\}$ represents each remote sensing image for a given period). In a similar vein, we use $D^{(l)}$ to symbolize the decoder.

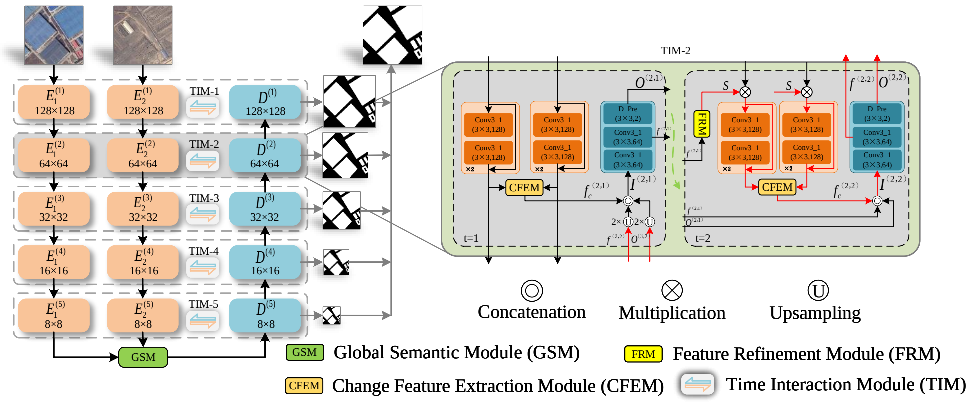

**Figure 1.** General structure of the bitemporal remote sensing image change detection network based on Siamese-attention feedback architecture (SAFNet). The entire network is an encoder–decoder structure, with $E_1^{(l)}$ and $E_2^{(l)}$ being the two decoder branches and $D^{(l)}$ being the encoder branch. The input dual-time remote sensing images first pass through $E_1^{(l)}$ and $E_2^{(l)}$ to extract multi-scale features, and then a global semantic module is established based on $E_1^{(5)}$ and $E_2^{(5)}$, outputting global semantic change information. The decoder network takes the global semantic change information and multi-scale features as inputs, generating increasingly significant change prediction maps layer by layer. We utilize the change feature extraction module (CFEM) to learn the difference information of dual-temporal remote sensing images. The feature refinement module (FRM) is used to control the information transfer at times t = 1 and t = 2.

Encoder network: In order to adapt ResNet34 for our change detection task, we modified it into a Siamese fully convolutional network, while removing the last two fully connected layers and the average pooling layer from the original network. In addition, we removed the downsizing operation of the last convolutional block $E_i^{(5)}$ with the goal of obtaining a larger range of features. We also used dilated convolution in this layer, as dilated convolution allows us to expand the receptive field without losing information and without requiring additional parameters. This equalizes the feature maps' sizes, $E_i^{(5)}$ and $E_i^{(4)}$, which is 1/16 of the initial input size. After each convolutional block, we obtain the difference information of each layer through the change feature extraction module (CFEM) so that the edge features and texture features can be better located during the decoding phase. The specific details are discussed in later sections.

Decoder Network: The decoder network consists of five convolutional blocks. In order to ensure that the output is of the same scale size as the encoder network, a two-fold upsampling is used between the decoder blocks, except for the $D^{(5)}$ layer. Each decoder block has three 3 × 3 convolutional layers, and there are 64, 64, and 2 output feature layers, correspondingly. The output channel is set to 64 in order to reduce the computational expense. During the process of feature interaction in the time interaction module (TIM), the feature information $f^{(l,1)}$ learned at time t = 1 is fed back to time t = 2 through the attention map $S^{(l)}$ generated by the feature refinement module (FRM). Detailed specifics will be addressed in subsequent chapters. During the network training process, each $D^{(l)}$ will cyclically generate two change prediction maps $O^{(l,1)}$ and $O^{(l,2)}$. Here, $O^{(l,1)}$ serves as the feedback from time t = 1 to time t = 2, and $O^{(l,2)}$ is the output prediction map of each $D^{(l)}$, and it is supervised with the real labels via the cross-entropy loss.

Global semantic module: We take advantage of the rich semantic features learned from $E_1^{(5)}$ and $E_2^{(5)}$ to generate low-resolution semantic change features. These features will be input into the decoder block layer by layer for refinement, and finally generate a change prediction map. This mechanism allows our model to extract meaningful information from the rich semantic features learned from both time-stamped images, and the global semantic module refines these features progressively to produce the change detection map.

### 2.1.2. Global Semantic Module (GSM)

We propose a global semantic module (GSM) because the single image features extracted directly by the Siamese network each have different characteristics and cannot be directly applied to the upsampling change map reconstruction process. We believe that the deepest pixels have the largest receptive field, which means that each pixel contains a large amount of feature information, and it is beneficial to fuse all feature information to determine semantic change information. Based on these facts, we propose a global semantic module (GSM) as shown in Figure 2. It fuses the two sets of features to fully utilize the global information, performs deep feature extraction on the fused features, and then recognizes the global semantic change information.

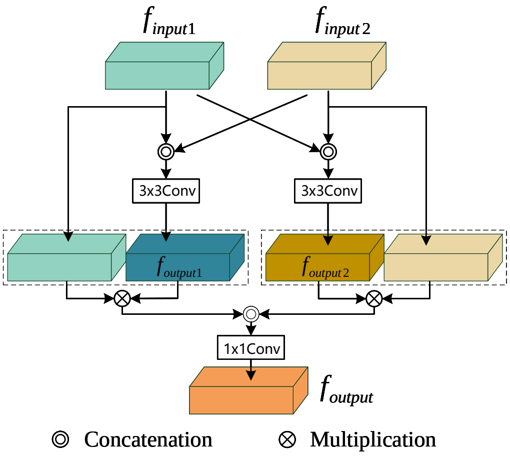

**Figure 2.** Global semantic module.

Suppose $f_{input1}$ and $f_{input2}$ are the two input features of the global semantic module, and the input size is $C \times H \times W$. First, to better detect the change in semantics, $f_{input1}$ and $f_{input2}$ are cross fused and dimensionally reduced through 3 × 3 convolution to obtain output features $f_{output1}$ and $f_{output2}$, and the size of the output feature map is $C \times H \times W$. Then, $f_{input1}$ and $f_{output1}$ are fused by weighting, and $f_{input2}$ and $f_{output2}$ are fused by weighting so that the fused features overcome the difficulties of different characteristics of a single image. This way, the bi-temporal semantic information extracted by the network is more representative and instructive, which enhances the network's ability to distinguish the change area. Finally, we perform a concatenation operation on the fused features and use 1 × 1 convolution for deep feature extraction to further extract semantic change information and refine it, obtaining the output feature map $f_{output}$, and the size of the output feature map is $C \times H \times W$. The following formula may be used to explain the aforesaid procedure' calculation:

$$f_{output1} = \sigma(f^{3\times3}([f_{input1}; f_{input2}])) \tag{1}$$

$$f_{output2} = \sigma(f^{3\times3}([f_{input2}; f_{input1}])) \tag{2}$$

$$f_{output} = f^{1\times1}([f_{input1} \otimes f_{output1}; f_{input2} \otimes f_{output2}]) \tag{3}$$

In the above, $\sigma(\cdot)$ represents the activation function *sigmoid*. $[;]$ represents the concatenation operation. $\otimes$ represents element-by-element multiplication. $f^{3\times3}(\cdot)$ and $f^{1\times1}(\cdot)$ are convolution layers having a convolution kernel of 3 and 1, respectively.

### 2.1.3. Change Feature Extraction Module (CFEM)

There are currently various methods for extracting change features, such as subtraction, addition, or concatenation operations, among others. However, the extraction of change features through these simple operations often hinders the network's discriminative ability. Particularly, the network's deep features, being abstract, are rendered unrecognizable by these operations, which leads to misjudgments and omissions during prediction. Hence, this section proposes a change feature extraction module (CFEM). It is depicted in Figure 3 as its structure. This module aims to extract the essential change features from the bitemporal remote sensing images to enhance the network's change detection discriminative ability.

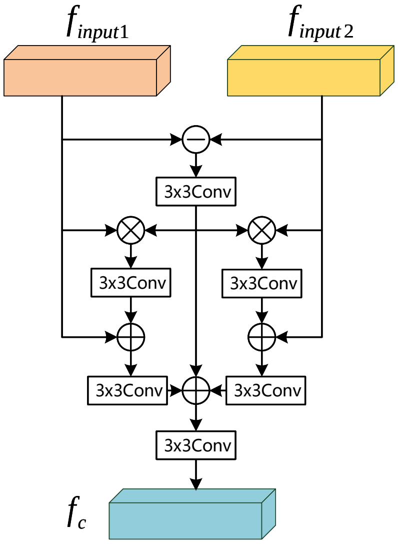

**Figure 3.** Change feature extraction module.

Assume that the inputs of the change feature extraction module (CFEM) are $f_{input1}$ and $f_{input2}$, the feature map size is $C \times H \times W$, where H and W represent the height and width of the feature map, and C denotes the number of channels of the feature map. Firstly, the input feature maps $f_{input1}$ and $f_{input2}$ are subtracted, and the absolute values are taken, yielding a difference feature map which then goes through a 3 × 3 convolution to extract multiscale difference feature information $f_{abs}$. To enable the network to adaptively select learning change region features and discard unnecessary information, thus reducing the redundancy of input features, the multiscale difference features $f_{abs}$ are weighted fused with the features of $f_{input1}$ and $f_{input2}$, respectively, and then go through an 3 × 3 convolution for deep feature learning. Afterwards, they are added and fused with $f_{input1}$ and $f_{input2}$, respectively, and then go through another 3 × 3 convolution to obtain $f_{output1}$ and $f_{output2}$. After that, $f_{abs}$, $f_{output1}$ and $f_{output2}$ are added together and fused, then go through another 3 × 3 convolution to finally obtain the output feature map $f_c$. The computational formula for the above process can be described as

$$f_{abs} = f^{3\times3}(abs(f_{input1} - f_{input2})) \tag{4}$$

$$f_{output1} = \sigma(f^{3\times3}(\sigma(f^{3\times3}(f_{abs} \otimes f_{input1})) + f_{input1})) \tag{5}$$

$$f_{output2} = \sigma(f^{3\times3}(\sigma(f^{3\times3}(f_{abs} \otimes f_{input2})) + f_{input2})) \tag{6}$$

$$f_c = f^{3\times3}(f_{abs} + f_{output1} + f_{output2}) \tag{7}$$

In this context, $abs(\cdot)$ represents the operation of taking the absolute value after subtraction.

### 2.1.4. Feature Refinement Module (FRM)

Considering that a single axial attention may lose information from another dimension, we introduce the feature refinement module (FRM) to fuse change feature information from both the channel and spatial dimensions. The structure of this module is shown in Figure 4. In current change detection tasks, the background of the change area is complex, and the target features are not clear, which may lead to omissions and misjudgments of the change area, hindering the final accuracy. Therefore, there is a need for feature refinement in both the spatial and channel dimensions, filtering noise information, suppressing non-change features, and highlighting change features. Our proposed feature refinement module (FRM) is well capable of achieving this. FRM extracts the changed features of the original input features in the channel dimension and space dimension, and then performs feature fusion with the original input features so that the network assigns higher weights to the changed areas and lower weights to the non-changed areas, and finally outputs more refined attention features.

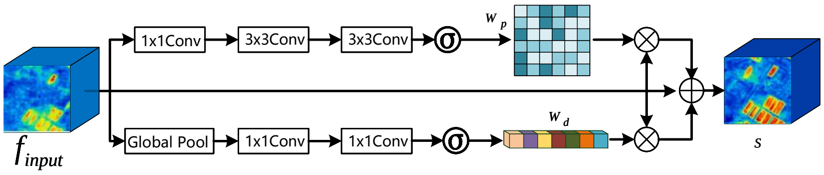

**Figure 4.** Feature refinement module.

Assume the size of the input feature map $f_{input}$ is $C \times H \times W$. Firstly, in the spatial dimension $f_{input}$, it goes through an 1 × 1 convolution, which compresses the feature channel number to 1, obtaining the $1 \times H \times W$ feature map, which then goes through two 3 × 3 convolution layers to extract feature information, followed by sigmoid activation to obtain the spatial weight coefficient matrix $W_p$. In the channel dimension $f_{input}$, it first goes through a global average pooling to obtain the $C \times 1 \times 1$ feature map, then goes through two 1 × 1 convolutions to extract channel characteristics, followed by *sigmoid* activation to obtain the channel weight coefficient matrix $W_d$. Afterwards, $W_p$ and $f_{input}$ are weighted and fused to obtain the spatial path output $f_p$, while $W_d$ and $f_{input}$ are weighted and fused to obtain the channel path output $f_d$. In the end of the feature refinement module (FRM), $f_{input}$, $f_p$ and $f_d$ are added and fused to obtain the output $S$. The mathematical expression for the process mentioned above can be stated as

$$f_p = \sigma(f^{3\times3}(f^{3\times3}(f^{1\times1}(f_{input})))) \otimes f_{input} \tag{8}$$

$$f_d = \sigma(f^{1\times1}(f^{1\times1}(AvgPool(f_{input})))) \otimes f_{input} \tag{9}$$

$$S = f_{input} + f_p + f_d \tag{10}$$

In this context, $AvgPool(\cdot)$ denotes the operation of global average pooling.

#### 2.1.5. Temporal Interaction Module (TIM)

Given the inherent characteristics of convolutional neural networks, if the preceding block fails to segment the entire target correctly, the erroneous predictions may accumulate in the subsequent blocks, resulting in an incomplete structure in the change detection results. Since we cannot guarantee the accuracy of feature information after refinement at the first time step, the guidance from the previous convolution block might involve magnification or reduction operations, bringing many inaccuracies, especially at the target boundaries. Moreover, if the preceding convolution block makes an error in learning and fails to detect the entire target, the subsequent blocks will not risk carrying out a structurally complete detection and identification. Therefore, we propose a temporal interaction module (TIM), the structure of which is shown in the right part of Figure 1. Considering that inaccurate guidance and overuse of some features may affect the blurring or drift of the detection target, in order to overcome this obstacle, in the temporal interaction module (TIM), we use FRM to control the information transfer and guide the network's learning process of boundary awareness and tiny targets through the feedback of t = 1 time to t = 2 time attention so that our network can not only integrate the features between different stages through guidance but also have the opportunity to correct errors at each stage through the attention feedback mechanism. Through TIM, feature feedback between two time steps is implemented to enhance the network's perception of changing targets and improve detection accuracy.

The TIM is used to control the information transfer between the decoder and encoder modules. The enlarged structure of this module is shown in the right part of Figure 1, and TIM works by cycling through two time steps. To clearly explain how TIM works, black and red lines are used to demonstrate the information flow across the two time steps. The output features from the change feature extraction module (CFEM) of $E_1^{(l)}$ and $E_2^{(l)}$ are marked as $f_c^{(l,t)}$, the output features of $D^{(l)}$ are marked as $f^{(l,t)}$, and the output prediction is marked as $O^{(l,t)}$, where t represents the time step. When $t$ = 1, the encoder block's $f_c^{(l,1)}$, as well as $f_c^{(l+1,2)}$ and $O^{(l+1,2)}$ from $D^{(l+1)}$, serve as the inputs for the decoder block $D^{(l)}$. To save memory, we use an $1 \times 1$ convolution on $f_c^{(l,1)}$ to reduce its channel number to 64, while performing $2\times$ upsampling on the output of $D^{(l+1)}$ to match the spatial size of $f_c^{(l,1)}$. Afterwards, we concatenate all the features in the channel dimension to form a rough prediction feature map $I^{(l,1)}$, which is input into the decoder block. The output feature and output prediction map of the first time step are $f^{(l,2)}$ and $O^{(l,2)}$ , respectively. The entire flow chart is shown on the right side of Figure 1 $(TIM - 2, t = 1)$. At t = 2, $f^{(l,1)}$ passes through the feature refinement module (FRM) to generate refined features S, providing guidance for deep feature learning at the t = 2 moment, and the encoder generates an updated difference feature map $f_c^{(l,2)}$. Then, $f_c^{(l,2)}$, $f^{(l,1)}$ and $O^{(l,1)}$ are concatenated to produce a new feature $I^{(l,2)}$. Finally, the decoder block refines the input features again, generating output features $f^{(l,2)}$ and output prediction maps $O^{(l,2)}$, which then proceed to the next stage. The entire flow chart is shown on the right side of Figure 1 $(TIM - 2, t = 2)$.

### *2.2. Datasets*

To more comprehensively validate the effectiveness of our proposed SAFNet model, we evaluate the model's performance on three different remote sensing image change detection datasets, BICD [50], CDD [51] and LEVIR-CD [43].

#### 2.2.1. BICD

The BICD dataset is a remote sensing image change detection dataset suggested in our first work. It includes 3420 pairs of bi-temporal remote sensing images with 512 × 512 pixels. The number of image pairs divided into the training set, validation set, and test set are 2280, 570, and 570, respectively. This dataset spans the period from 2010 to 2019, and includes paired images from different regions of Eastern China at different times. The objects in the images include factories, farmlands, roads, and buildings, among other areas. The schematic diagram of the dataset is shown in Figure 5.

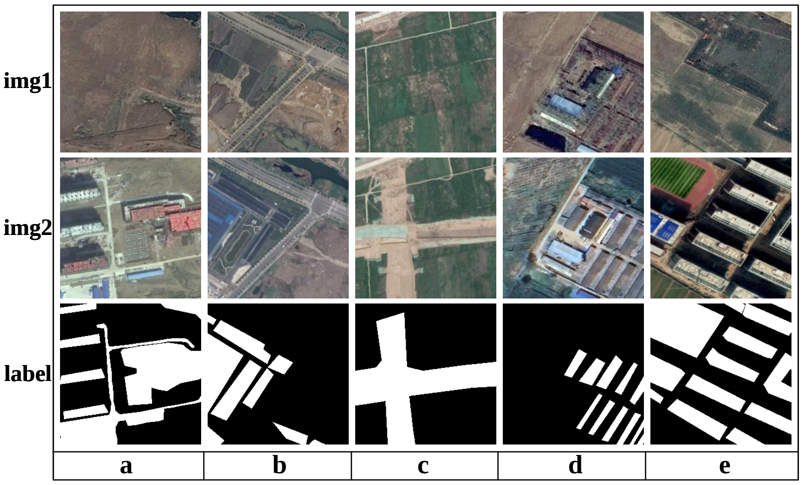

**Figure 5.** BICD schematic diagram. A few dual-temporal remote sensing images from the dataset are displayed, with each column constituting a sample. The first and the second rows depict the dual-temporal remote sensing images, and the third row shows the labels (with white representing areas of change, and black signifying unchanged areas). (**a**–**e**) are 5 pairs of dual-temporal remote sensing image pairs and corresponding labels selected from BICD.

### 2.2.2. CDD

The CDD dataset is a publicly available remote sensing image change detection dataset. The original images of this dataset are composed of seven pairs of images with a resolution of 4750 × 2700 and four pairs of images with a resolution of 1900 × 1000. These 11 pairs of images are synchronously cropped into 16,000 pairs of 256 × 256 pixel images, with the number of image pairs divided into the training set, validation set, and test set being 10,000, 3000, and 3000, respectively. The schematic diagram of this dataset is shown in Figure 6.

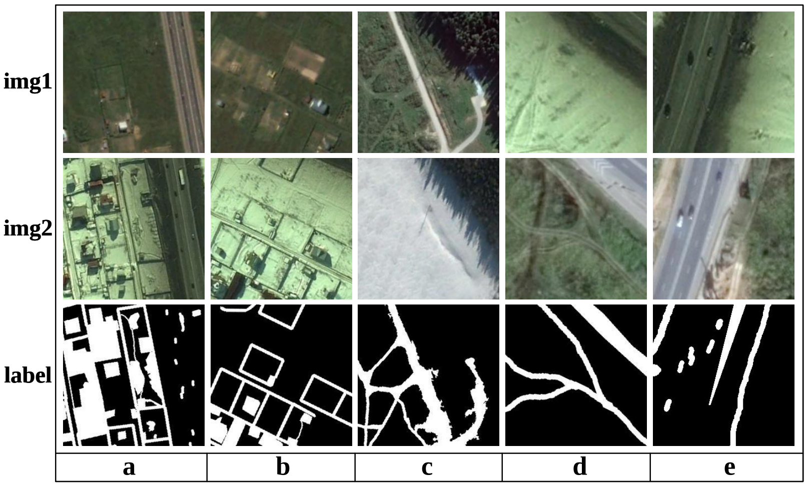

**Figure 6.** CDD dataset schematic diagram. Each column represents one sample, with the first and second rows displaying the bi-temporal remote sensing images, while the third row shows the labels (white denotes change areas, black denotes no-change areas). (**a**–**e**) are 5 pairs of dual-temporal remote sensing image pairs and corresponding labels selected from CDD.

#### 2.2.3. LEVIR-CD

LEVIR-CD is a large-scale building change detection dataset consisting of 637 pairs of 1024 × 1024 ultra-high resolution remote sensing images. It has a time span of 5–14 years, filmed between 2002 and 2018, focusing on major changes in architecture. The dual-temporal remote sensing images in the dataset come from 20 different areas in multiple cities in Texas, including villas, high-rise apartments, small garages, large warehouses, and other buildings, and consider seasonal and illumination changes. We crop the images into 256 × 256 pixel images, and the image logarithms are 7120, 1024 and 2048 for training set, validation set and test set, respectively. Its schematic diagram is shown in Figure 7.

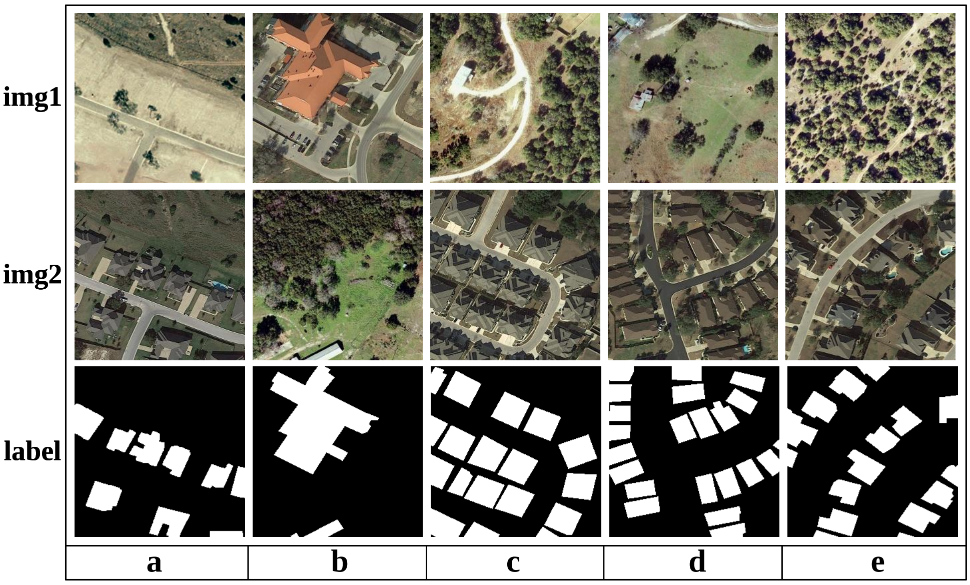

**Figure 7.** LEVIR-CD dataset schematic diagram. Each column represents one sample, with the first and second rows displaying the bi-temporal remote sensing images, while the third row shows the labels (white denotes change areas, black denotes no-change areas). (**a**–**e**) are 5 pairs of dual-temporal remote sensing image pairs and corresponding labels selected from LEVIR-CD.

### 2.3. Implementation Details

#### 2.3.1. Evaluation Metrics

To assess the performance of our SAFNet model in change detection tasks, we adopted five commonly used evaluation metrics: precision (PR), recall (RC), mean intersection over union (MIoU), F1-score (F1) and pixel accuracy (PA). We use F1 of the changed category and MIoU of the changed category and unchanged category as the main evaluation indicators, and PR, RC and PA as auxiliary indicators to comprehensively evaluate the model:

$$PR = \frac{TP}{TP + FP} \tag{11}$$

$$RC = \frac{TP}{TP + FN} \tag{12}$$

$$MIOU = \frac{TP}{TP + FP + FN} \tag{13}$$

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$F1 = \frac{2 \times PR \times RC}{PR + RC} \tag{15}$$

Here, $TP$ representing true positive indicates the correct prediction of changing areas; $TN$ representing true negative refers to the correct prediction of non-changing areas; $FP$ representing false positive implies predicting non-changing areas as changing areas; and $FN$ representing false negative denotes predicting changing areas as non-changing areas.

### 2.3.2. Experimental Details

The experiments in this paper are implemented on a GeForce RTX 3080 graphics processing unit (GPU) based on PyTorch. We set the batch size for training to 6 and the initial learning rate $lr$ to 0.001, and dynamically adjust the learning rate using a ploy strategy. he mathematical formula for its calculation is as follows:

$$lr = lr\_ * (1 - \frac{epoch}{num\_epoch})^{p} \tag{16}$$

Here, $lr$ is the new learning rate, $lr\_$ is the initial learning rate, epoch is the current iteration number, $num\_epoch$ is the maximum iteration number, and $p$ is a constant that manages the speed of decay. The epoch is set to 200. Furthermore, we choose binary cross entropy loss as the loss function for our network, and use Adam as the optimizer for our network.

## 3. Results

In this part, in order to comprehensively evaluate the effectiveness of our proposed network and modules, we conduct ablation experiments and comparative experiments on BICD, CDD and LEVIR-CD datasets.

### *3.1. Network Structure Selection*

We compare the experimental results of two CNN architectures, the early fusion architecture and the Siamese architecture. Early fusion network structures concatenate two images together before feeding them into the network, treating them as different color channels. The Siamese structure processes two images separately through a network with the same structure and weight parameters. The experimental results in Table 1 show that the Siamese structure outperforms the early fusion structure, and the F1 and MIoU are improved by 0.95% and 0.74%, respectively.

**Table 1.** Performance comparison of different network structures (the best results are indicated in bold font).

| Method | PA (%) | PR (%) | RC (%) | F1 (%) | MIoU (%) |
|---|---|---|---|---|---|
| Early Fusion | 95.72 | 88.15 | 78.64 | 82.57 | 83.14 |
| Siamese | **95.89** | **89.75** | **79.12** | **84.09** | **83.31** |

### *3.2. Ablation Experiments on BICD*

Given the complexity of our network, we can gain a better understanding of it and verify the effectiveness of each module by either deleting or adding modules that we propose. Thus, we conduct ablation experiments on each module using the BICD dataset. Table 2 shows our experimental results, and all models have the same training strategy. To intuitively compare the effectiveness of our models, we primarily focus on the MIoU metric. In the experiments, we use ResNet34 as the backbone of our network.

(1) Ablation experiment of CSM: By introducing CSM into the deep layer of the decoder network and fully utilizing global information to capture change information, we generate a low-resolution semantic change map. This map can guide the repair of shallow texture information and help the network reduce missed detections and false detections during the decoding phase. The experimental results in Table 2 show that GSM improves the scores of F1 and MIoU by 0.73% and 1.79% over the backbone network, proving the effectiveness of GSM.

(2) Ablation experiment of TIM: In order to reduce the accumulation of incorrect predictions in subsequent blocks, our proposed TIM can enable feature feedback between two time steps to enhance the network's perception of changing targets, and to repair the details of changing areas, thereby increasing the accuracy of feature learning. The experimental results in Table 2 show that our proposed TIM improves the scores of F1 and MIoU by 0.22% and 0.36%.

(3) Ablation experiment of CFEM: To extract important change features from bi-temporal remote sensing images, our proposed CFEM performs change feature extraction via absolute difference operations, improving the learning of edge features and texture features of changing areas, thus enhancing the network's discriminative ability. The experimental results in Table 2 show that our proposed CFEM improves the scores of F1 and MIoU by 0.27% and 0.44%, demonstrating the effectiveness of our proposed module.

(4) Ablation experiment of FRM: To fuse channel information and spatial information, our proposed FRM enables the network to simultaneously capture change features in both spatial and channel dimensions, removes and suppresses redundant features, and then weights feature extraction for the next time step, thereby improving the network's detection accuracy. The experimental results in Table 2 show that our proposed FRM improves the scores of F1 and MIoU by 0.19% and 0.33%, demonstrating the effectiveness of our proposed module. At the same time, we use Figure 8 to illustrate the effectiveness of the FRM module more intuitively. The figure shows the heat map effect of adding FRM and not adding FRM. It can be seen that after the introduction of FRM, the network effectively solves the problem of fuzzy edge details of changing targets and the false detection of tiny targets.

**Table 2.** The effectiveness of our suggested module is assessed by ablative experiments (the best results are indicated in bold font).

| Method | PA (%) | PR (%) | RC (%) | F1 (%) | MIoU (%) |
|---|---|---|---|---|---|
| Backbone | 95.09 | 88.54 | 77.73 | 82.78 | 80.39 |
| Backbone + GSM | 95.59 | 88.87 | 78.75 | 83.51 | 82.18 |
| Backbone + GSM + TIM | 95.69 | 89.57 | 78.62 | 83.73 | 82.54 |
| Backbone + GSM + TIM + CFEM | 95.81 | 89.48 | 78.99 | 83.90 | 82.98 |
| Backbone + GSM + TIM + CFEM + FRM | **95.89** | **89.75** | **79.12** | **84.09** | **83.31** |

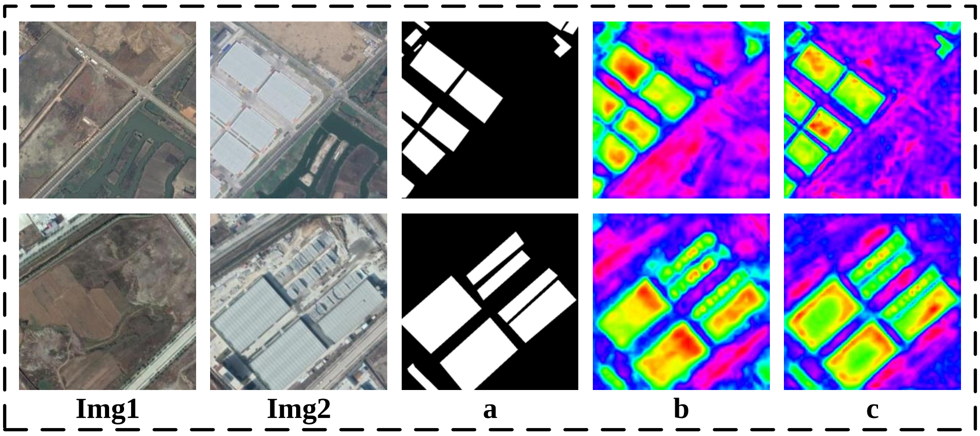

**Figure 8.** Heatmaps without and with FRM are compared. (**Img1**,**Img2**) are dual-temporal remote sensing images. (**a**) labels, (**b**) feature heatmap without FRM module in the network, (**c**) feature heatmap with FRM module added to the network.

The message passing between the decoder and encoder modules is controlled by a temporal interaction module (TIM), which is the core innovation of our approach. The module works in a two-time-step loop. In the first time step, the weight-shared Siamese network extracts the change features of the image through CFEM, and the output features and output prediction maps extracted in the second time step of the previous stage are sent to the decoding module. At this time, after the feature extraction of the first time step, we cannot guarantee the quality of the result because the features from the previous block will bring many uncertain values to the boundary of the object after the enlargement operation, and if for the previous block, one module fails to identify a small variation target, subsequent modules will not risk performing structurally complete detection. Therefore, we use feature extraction in two time steps to enhance the learning ability of the network for edge and tiny targets, introduce FRM between the two time steps, refine and filter the features learned in the first time step, highlight the changing features to suppress non-changing features (the results of Figure 8 show that the effect of this module is significant, used to guide the feature extraction of the second time step), and finally generate a salient prediction map with finer boundaries and more prominent small targets. In order to verify the effectiveness of TIM with two time steps, we conduct an ablation comparison experiment with TIM with one time step. The experimental results are shown in Table 3. TIM_(t = 1) means having one time step, and TIM_(t = 1, t = 2) means having two time steps. As can be seen from the table, TIM_(t = 1, t = 2) works better.

**Table 3.** TIM ablation comparison experiment (the best results are indicated in bold font).

| Method | PA (%) | PR (%) | RC (%) | F1 (%) | MIoU (%) |
|---|---|---|---|---|---|
| TIM_(t = 1) | 95.55 | 88.63 | 79.01 | 83.54 | 82.48 |
| TIM_(t = 1,t = 2) | **95.89** | **89.75** | **79.12** | **84.09** | **83.31** |

Our proposed modules, including the global semantic module (GSM), time interaction module (TIM), change feature extraction module (CFEM), and feature refinement module (FRM), further optimize the performance of the network based on the backbone network. By individually introducing these modules, the scores of the evaluation index MIoU are increased by 1.79%, 0.36%, 0.44% and 0.33%, respectively, and the scores of F1 are increased by 0.73%, 0.22%, 0.27% and 0.19%. And through the four modules, for the joint effect of our model, compared with the backbone network, the score of MIoU is improved by 2.92%, and the score of F1 is improved by 1.41%. These four modules can thereby enhance the learning capacity and prediction accuracy of the network for change detection tasks.

### *3.3. Comparison Experiments with Different Algorithms on BICD*

In this section, we compare the method we propose with other semantic segmentation and change detection methods to observe our model comprehensively. In the experiments, to ensure fairness in the comparison, all methods use the same training strategy. The quantitative results of various model metrics are shown in Table 4. From the table, we can see that the traditional change detection method PCA-Means has very poor detection effects and can hardly complete the change detection task on this dataset. Among the deep learning methods, FC-Siam-Diff has the worst effect with scores of F1 and MIoU of only 57.97% and 64.51%, and the detection effect of SAGNet is the best compared with other deep learning methods, with the scores of F1 and MIoU reaching 83.45% and 83.93%. Compared with other algorithms, our model SAFNet basically achieves the best results according to various indicators; its MIoU is slightly lower than that of SAGNet, but its F1 and MIoU also reach 84.09% and 83.31%, indicating the effectiveness of our proposed method.

Figure 9 shows the comparison of prediction maps of different methods. We compare prediction maps on three different sets of bi-temporal remote sensing images to further illustrate the feasibility of our algorithm. Among them, $a$ is the image label, and $b - l$ are prediction maps from various comparative algorithms. A comprehensive review of the three sets of comparative experiments reveals that the traditional PCA-Means algorithm performs the worst. Next is the FC series change detection algorithm, which can hardly recognize change areas and suffers from many missed detections and false alarms. The prediction maps of other deep learning-based algorithms are rather coarse, and their abilities to handle edge details and small targets are inferior. In contrast, the prediction map generated by our algorithm predicts the edge details of the change area and the changes in small targets very well. It can be said to differentiate between change and non-change areas in a targeted manner. Moreover, our algorithm's prediction map is closer to the label and yields better prediction results.

**Table 4.** Results of experiments compared to those from other algorithms (the best results are indicated in bold font).

| Method | PA (%) | PR (%) | RC (%) | F1 (%) | MIoU (%) |
|---|---|---|---|---|---|
| PCA-Means [31] | 86.32 | 28.66 | 12.53 | 17.43 | 48.82 |
| FC-Siam-Diff [53] | 90.79 | 77.63 | 46.26 | 57.97 | 64.51 |
| FC-EF | 90.23 | 73.26 | 43.24 | 54.37 | 65.94 |
| FC-Siam-Conc | 91.45 | 78.18 | 47.55 | 59.14 | 68.66 |
| Unet [54] | 92.57 | 81.36 | 69.73 | 75.09 | 72.59 |
| FCN-8s [55] | 93.02 | 81.67 | 72.75 | 76.95 | 74.38 |
| ChangNet [56] | 94.14 | 88.57 | 77.11 | 82.44 | 76.49 |
| DASNet [57] | 94.84 | 89.34 | 75.46 | 81.82 | 79.93 |
| TCD-Net [58] | 95.24 | 88.28 | 74.03 | 80.53 | 81.13 |
| MFGAN [59] | 95.44 | 87.99 | 76.18 | 81.66 | 82.09 |
| BIT [60] | 95.78 | 89.51 | 75.68 | 82.02 | 82.89 |
| TFI-GR [61] | 95.63 | 88.87 | 77.92 | 83.04 | 82.96 |
| SAGNet | 95.82 | 89.36 | 78.28 | 83.45 | **83.93** |
| SAFNet (our) | **95.89** | **89.75** | **79.12** | **84.09** | 83.31 |

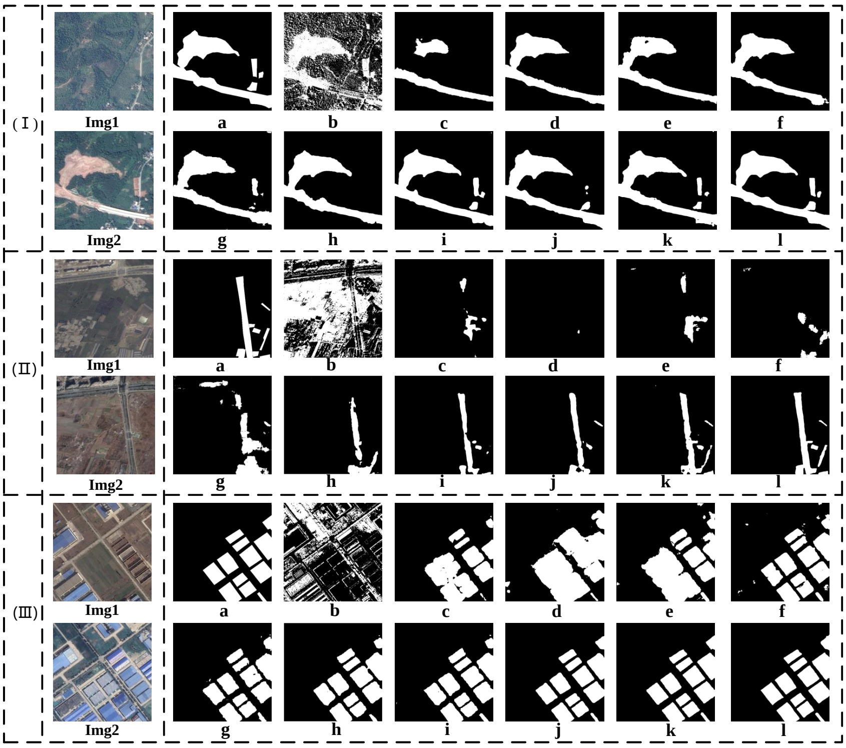

**Figure 9.** Presents the contrast of predictive results from various algorithms. (**I–III**) correspond to the comparative experiments for three sets of bi-temporal remote sensing images. (**Img1**,**Img2**) are indicative of remote sensing images captured at different time intervals. (**a**–**l**) respectively represent prediction maps of the label, PCA-Means, FC-Siam-Diff, FC-EF, FC-Siam-Conc, Unet, FCN-8s, ChangNet, DASNet, TCD-Net, MFGAN, and our network SAFNet.

### *3.4. Generalization Experiments on the CDD Dataset*

To further verify the effectiveness of our SAFNet algorithm, we conducted generalization experiments on the CDD dataset, with the results shown in Table 5. From the F1 and MIoU scores, it can be seen that the PCA-Means method is the lowest, and the F1 and MIoU scores are only 31.23% and 39.53%. The rest of the algorithms based on deep learning are improved in the four indicators, and the best performance is our algorithm. SAFNet, F1 and MIoU scores reach 89.48% and 73.36%, which shows that our algorithm has very good generalization and robust performance.

**Table 5.** Comparison of generalization experimental results on the CDD dataset (the best results are indicated in bold font).

| Method | PA (%) | PR (%) | RC (%) | F1 (%) | MIoU (%) |
|---|---|---|---|---|---|
| PCA-Means | 80.35 | 39.36 | 25.88 | 31.23 | 39.53 |
| FC-EF | 94.39 | 85.31 | 59.86 | 70.35 | 52.38 |
| FC-Siam-Diff | 94.92 | 84.32 | 63.51 | 72.45 | 53.27 |
| FC-Siam-Conc | 94.78 | 83.69 | 64.32 | 72.74 | 53.88 |
| FCN-8s | 97.01 | 83.13 | 75.06 | 78.89 | 68.19 |
| Unet | 97.66 | 84.24 | 74.57 | 79.11 | 68.83 |
| DASNet | 97.47 | 84.85 | 89.79 | 87.25 | 70.21 |
| ChangNet | 97.64 | 82.27 | 90.21 | 86.07 | 70.93 |
| TCD-Net | 97.39 | 83.65 | 91.32 | 87.32 | 71.72 |
| MFGAN | 97.37 | 83.76 | 92.83 | 88.05 | 72.21 |
| TFI-GR | 97.58 | 84.53 | 92.63 | 88.41 | 72.39 |
| BIT | 97.49 | 83.57 | 93.88 | 88.43 | 73.01 |
| SAFNet(our) | **97.67** | **85.32** | **94.06** | **89.48** | **73.36** |

Figure 10 shows the prediction comparison diagram of different algorithms on the CDD dataset. (I), (II), and (III) are three groups of different bi-temporal remote sensing images, and $b - l$ are prediction maps of various algorithms. As can be seen from the figure, compared to other methods, our proposed SAFNet algorithm predicts edge details and small targets more accurately, and the predicted change areas are also clearer. Other methods all have some shortcomings. Especially for image (I), the change area is not obvious, and the interference from the background is strong. Under such circumstances, other methods can only predict part of the change area, some even cannot predict at all, while our method can accurately predict continuous change areas, overcome background interference, and handle edge parts more delicately. This demonstrates the effectiveness of our SAFNet algorithm for change detection tasks.

### *3.5. Generalization Experiments on the LEVIR-CD Dataset*

The generalization results for LEVIR-CD are shown in Table 6. The results of the generalization experiment show that among other change detection methods, TFI-GR obtained the best detection results, and the scores of F1 and MIoU reached 90.01% and 89.37%. However, our SAFNet still achieves the highest accuracy on LEVIR-CD, the scores of F1 and MIoU reach 90.67% and 89.66%, and F1 and MIoU are improved by 0.66% and 0.29% compared with TFI-GR.

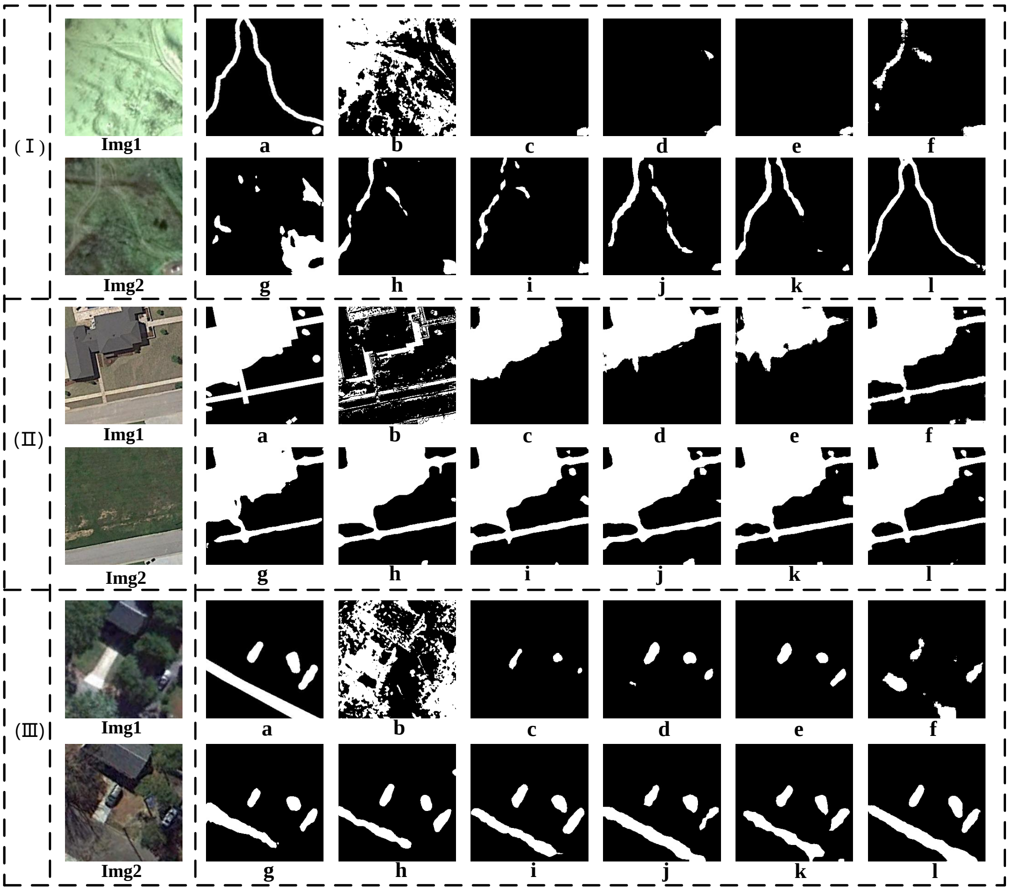

**Figure 10.** Presents the contrast of predictive results from various algorithms. (**I–III**) correspond to the comparative experiments for three sets of bi-temporal remote sensing images. (**Img1**,**Img2**) are indicative of remote sensing images captured at different time intervals. (**a**–**l**) correspond to the prediction maps of the label, PCA-Means, FC-EF, FC-Siam-Diff, FC-Siam-Conc, FCN-8s, Unet, DASNet, ChangNet, TCD-Net, MFGAN, and our SAFNet network, respectively.

**Table 6.** Comparison of generalization experimental results on the LEVIR-CD dataset (the best results are indicated in bold font).

| Method | PA (%) | PR (%) | RC (%) | F1 (%) | MIoU (%) |
|---|---|---|---|---|---|
| PCA-Means | 78.63 | 12.34 | 45.69 | 19.43 | 33.96 |
| DASNet | 98.57 | 90.35 | 84.23 | 87.19 | 87.43 |
| ChangNet | 98.48 | 90.54 | 86.98 | 88.72 | 86.64 |
| BIT | 98.62 | 90.26 | 88.51 | 89.38 | 89.19 |
| TFI-GR | 98.68 | 92.01 | 88.08 | 90.01 | 89.37 |
| SAFNet(our) | **98.87** | **92.49** | **88.93** | **90.67** | **89.66** |

Figure 11 is a comparison of the predictions of different algorithms on the LEVIR-CD dataset. Among them, img1 and img2 are a pair of dual-temporal remote sensing images to be detected. a represents the label, while b-g are the prediction results of different change detection algorithms. It can be seen from the figure that the edge details in the prediction images of other algorithms are very poor, and the prediction images of some algorithms even have serious missed and false detections. However, compared with other algorithms, the prediction map of SAFNet proposed by us handles the edge part well, and there is no missed detection and false detection phenomenon, which shows the superiority of our proposed SAFNet algorithm.

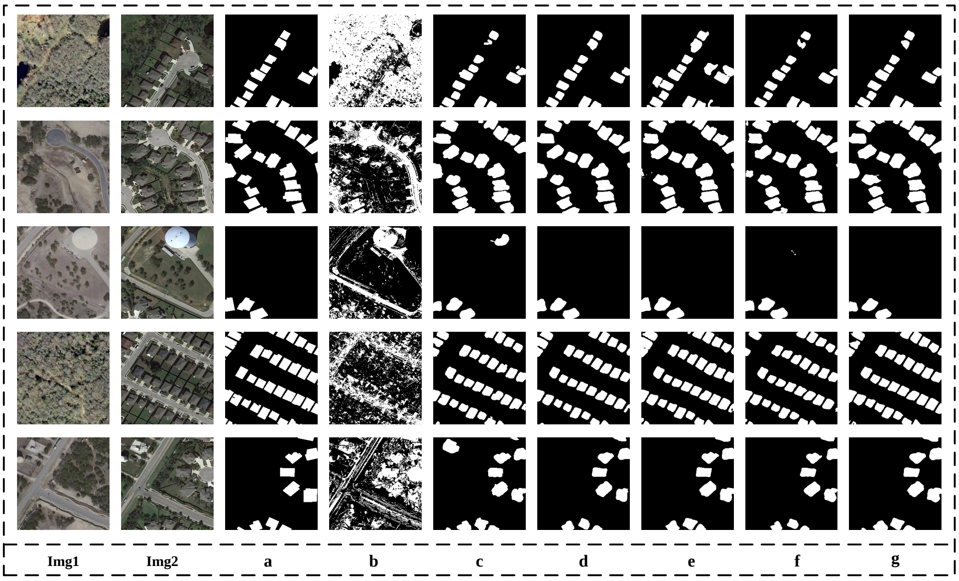

**Figure 11.** Comparison of prediction results graphs of different algorithms. (**Img1**,**Img2**) represent remote sensing images in different periods. (**a**–**g**) represents the prediction graph of label, PCA-Means, DASNet, ChangNet, BIT, Segformer, and our network SAFNet, respectively.

## 4. Discussion

### *4.1. Advantages of the Proposed Method*

It is proved by experiments that the method proposed in this paper is obviously superior to other algorithms, and can effectively detect the change area in the dual-temporal remote sensing image, and the experimental results on the three datasets prove the effectiveness and superiority of SAFNet. Compared with other methods, our method possesses higher detection accuracy. In the feature encoding stage, features are extracted through a weight-shared Siamese network and input into CFEM to generate rough change features; however, only combining the output of the decoder is not enough to detect changes in remote sensing images. Considering that the features from the previous block will bring many uncertain values to the boundary of the object after the upscaling operation, in addition, if the previous module fails to identify the small variation objects, the subsequent modules will not risk performing structurally complete detection. We perform cyclic learning on two time steps through TIM, and perform attention optimization through FRM between two time steps to adjust the weight between each channel and pixel so that the network pays more attention to the changing area and more finely learns edge features. In order to enable the network to detect false detections and missed detections of tiny objects, we introduce CSM at the bottom layer of the decoder network to learn features of global information, generate a low-resolution semantic change map, and guide the network to repair and optimize shallow texture information. Reduce the phenomenon of missed detection and false detection. All in all, our proposed method has achieved good results on different types of remote sensing images, and this method has strong generalization and robustness.

### *4.2. Limitations and Future Research Directions*

In our study, we introduce a Siamese-attention feedback architecture-based change detection network (SAFNet) for bitemporal remote sensing images, which has shown remarkable results on photo-level images. However, we also recognize that remote sensing images are often acquired through different means, such as synthetic aperture sonar (SAS) and synthetic aperture radar (SAR) [62,63]. The characteristics of these images, such as noise type, lighting conditions, etc., may differ from photorealistic images, which may have an impact on the performance of our method. We expect that SAFNet should work well for SAS and SAR images as well, although some adjustments may be needed to accommodate the characteristics of such images. For example, we may need to perform specific preprocessing on input images, or tune our network architecture and parameters to better handle the noise and lighting conditions of such images. Future research directions may include more experiments on SAS and SAR images to evaluate and optimize the performance of our method on such images. We also expect to further investigate possible improvements and optimizations to make our method better serve various fields of change detection in remote sensing images.

## 5. Conclusions

In this paper, we propose a bitemporal remote sensing image change detection network based on Siamese-attention feedback architecture (SAFNet). The network adopts an overall decoder–encoder structure. Through the time interaction module (TIM) established between the decoder and the encoder, the network can enhance the interaction of feature information between two time steps, filter out redundant information, and improve the network's perception of the entire change target. By introducing the global semantic module (GSM) in the encoder network, the network can capture the semantic change map of the change target, providing guidance for the reconstruction of the prediction map. In addition, we propose two modules: change feature extraction module (CFEM) and feature refinement module (FRM). CFEM captures multi-scale temporal difference information, enabling the network to learn edge features and texture features of the change area better. FRM focuses adaptively on the change area, improving the network's detection capabilities for edge details and small targets. Experimental results show that our proposed SAFNet outperforms other algorithms on BICD, CDD and LEVIR-CD datasets, and the F1 and MIoU indicators reach 84.09%, 89.48%, 90.67% and 83.31% and 76.36%, 89.66%, with very good generalization and robustness.

**Author Contributions:** Conceptualization, H.Y. and L.W.; methodology, M.X. and H.Y.; software, H.Y. and C.M.; validation, L.W. and H.L.; formal analysis, H.L.; investigation, H.Y. and C.M.; resources, M.X.; data curation, H.Y.; writing—original draft preparation, H.Y. and C.M.; writing—review and editing, M.X.; visualization, L.W.; supervision, L.W.; project administration, M.X.; funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported in part by the National Natural Science Foundation of PR China (42075130).

**Data Availability Statement:** The data and the code of this study are available from the corresponding author upon request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Zhang, Z.; Liu, F.; Zhao, X.; Wang, X.; Shi, L.; Xu, J.; Yu, S.; Wen, Q.; Zuo, L.; Yi, L.; et al. Urban expansion in China based on remote sensing technology: A review. *Chin. Geogr. Sci.* **2018**, *28*, 727–743.
2. Albalawi, E.K.; Kumar, L. Using remote sensing technology to detect, model and map desertification: A review. *J. Food Agric. Environ.* **2013**, *11*, 791–797.
3. Zhao, S.; Wang, Q.; Li, Y.; Liu, S.; Wang, Z.; Zhu, L.; Wang, Z. An overview of satellite remote sensing technology used in China's environmental protection. *Earth Sci. Inform.* **2017**, *10*, 137–148. [CrossRef]
4. Tong, Q.; Xue, Y.; Zhang, L. Progress in hyperspectral remote sensing science and technology in China over the past three decades. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *7*, 70–91. [CrossRef]
5. Weng, L.; Pang, K.; Xia, M.; Lin, H.; Qian, M.; Zhu, C. Sgformer: A Local and Global Features Coupling Network for Semantic Segmentation of Land Cover. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6812–6824. [CrossRef]
6. Chen, K.; Xia, M.; Lin, H.; Qian, M. Multi-scale Attention Feature Aggregation Network for Cloud and Cloud Shadow Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3283435. [CrossRef]
7. Dai, X.; Xia, M.; Weng, L.; Hu, K.; Lin, H.; Qian, M. Multi-Scale Location Attention Network for Building and Water Segmentation of Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3276703. [CrossRef]
8. Ji, H.; Xia, M.; Zhang, D.; Lin, H. Multi-Supervised Feature Fusion Attention Network for Clouds and Shadows Detection. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 247. [CrossRef]

9. Chen, J.; Xia, M.; Wang, D.; Lin, H. Double Branch Parallel Network for Segmentation of Buildings and Waters in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1536. [CrossRef]
10. Song, L.; Xia, M.; Weng, L.; Lin, H.; Qian, M.; Chen, B. Axial cross attention meets CNN: Bibranch fusion network for change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 32–43. [CrossRef]
11. Wang, D.; Weng, L.; Xia, M.; Lin, H. MBCNet: Multi-Branch Collaborative Change-Detection Network Based on Siamese Structure. *Remote Sens.* **2023**, *15*, 2237. [CrossRef]
12. Ma, C.; Weng, L.; Xia, M.; Lin, H.; Qian, M.; Zhang, Y. Dual-branch network for change detection of remote sensing image. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106324. [CrossRef]
13. Wing, M.G.; Burnett, J.D.; Sessions, J. Remote sensing and unmanned aerial system technology for monitoring and quantifying forest fire impacts. *Int. J. Remote Sens. Appl.* **2014**, *4*, 18–35. [CrossRef]
14. Qu, Y.; Xia, M.; Zhang, Y. Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow. *Comput. Geosci.* **2021**, *157*, 104940. [CrossRef]
15. Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; Lin, H. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote Sens.* **2022**, *43*, 5940–5960. [CrossRef]
16. Koltunov, A.; Ustin, S. Early fire detection using non-linear multitemporal prediction of thermal imagery. *Remote Sens. Environ.* **2007**, *110*, 18–28. [CrossRef]
17. Chen, B.; Xia, M.; Qian, M.; Huang, J. MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images. *Int. J. Remote Sens.* **2022**, *43*, 5874–5894. [CrossRef]
18. Ma, Z.; Xia, M.; Weng, L.; Lin, H. Local Feature Search Network for Building and Water Segmentation of Remote Sensing Image. *Sustainability* **2023**, *15*, 3034. [CrossRef]
19. Xian, G.; Homer, C. Updating the 2001 National Land Cover Database impervious surface products to 2006 using Landsat imagery change detection methods. *Remote Sens. Environ.* **2010**, *114*, 1676–1686. [CrossRef]
20. Hu, K.; Li, M.; Xia, M.; Lin, H. Multi-Scale Feature Aggregation Network for Water Area Segmentation. *Remote Sens.* **2022**, *14*, 206. [CrossRef]
21. Torres-Vera, M.; Prol-Ledesma, R.; García-López, D. Three decades of land use variations in Mexico City. *Int. J. Remote Sensinginternational J. Remote Sens.* **2009**, *30*, 117–138. [CrossRef]
22. Ma, Z.; Xia, M.; Lin, H.; Qian, M.; Zhang, Y. FENet: Feature enhancement network for land cover classification. *Int. J. Remote Sens.* **2023**, *44*, 1702–1725. [CrossRef]
23. Hu, K.; Zhang, E.; Xia, M.; Weng, L.; Lin, H. Mcanet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 1055. [CrossRef]
24. Lu, C.; Xia, M.; Qian, M.; Chen, B. Dual-branch network for cloud and cloud shadow segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
25. Gao, J.; Weng, L.; Xia, M.; Lin, H. MLNet: Multichannel feature fusion lozenge network for land segmentation. *J. Appl. Remote Sens.* **2022**, *16*, 16513.
26. Hu, K.; Zhang, D.; Xia, M.; Qian, M.; Chen, B. LCDNet: Light-weighted cloud detection network for high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4809–4823. [CrossRef]
27. Ke, L.; Lin, Y.; Zeng, Z.; Zhang, L.; Meng, L. Adaptive change detection with significance test. *IEEE Access* **2018**, *6*, 27442–27450. [CrossRef]
28. Rignot, E.J.; Van Zyl, J.J. Change detection techniques for ERS-1 SAR data. *IEEE Trans. Geosci. Remote Sens.* **1993**, *31*, 896–906. [CrossRef]
29. Kuncheva, L.I.; Faithfull, W.J. PCA feature extraction for change detection in multidimensional unlabeled data. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *25*, 69–80. [CrossRef]
30. Deng, J.; Wang, K.; Deng, Y.; Qi, G. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *Int. J. Remote Sens.* **2008**, *29*, 4823–4838. [CrossRef]
31. Celik, T. Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [CrossRef]
32. Zhang, X.; Cui, J.; Wang, W.; Lin, C. A study for texture feature extraction of high-resolution satellite images based on a direction measure and gray level co-occurrence matrix fusion algorithm. *Sensors* **2017**, *17*, 1474. [CrossRef] [PubMed]
33. Guiming, S.; Jidong, S. Remote sensing image edge-detection based on improved Canny operator. In Proceedings of the 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), Beijing, China, 4–6 June 2016; pp. 652–656.
34. Peng, D.; Zhang, Y.; Guan, H. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
35. He, P.; Shi, W.; Zhang, H.; Hao, M. A novel dynamic threshold method for unsupervised change detection from remotely sensed images. *Remote Sens. Lett.* **2014**, *5*, 396–403. [CrossRef]
36. Thonfeld, F.; Feilhauer, H.; Braun, M.; Menz, G. Robust Change Vector Analysis (RCVA) for multi-sensor very high resolution optical satellite data. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *50*, 131–140. [CrossRef]
37. Zheng, Y.; Zhang, X.; Hou, B.; Liu, G. Using combined difference image and *k*-means clustering for SAR image change detection. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 691–695. [CrossRef]

38. Luppino, L.T.; Bianchi, F.M.; Moser, G.; Anfinsen, S.N. Unsupervised image regression for heterogeneous change detection. *arXiv* **2019**, arXiv:1909.05948.
39. Zhang, C.; Weng, L.; Ding, L.; Xia, M.; Lin, H. CRSNet: Cloud and Cloud Shadow Refinement Segmentation Networks for Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 1664. [CrossRef]
40. Gong, M.; Zhao, J.; Liu, J.; Miao, Q.; Jiao, L. Change detection in synthetic aperture radar images based on deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 125–138. [CrossRef]
41. Zhan, Y.; Fu, K.; Yan, M.; Sun, X.; Wang, H.; Qiu, X. Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1845–1849. [CrossRef]
42. Zhang, M.; Xu, G.; Chen, K.; Yan, M.; Sun, X. Triplet-based semantic relation learning for aerial remote sensing image change detection. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 266–270. [CrossRef]
43. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
44. Song, L.; Xia, M.; Jin, J.; Qian, M.; Zhang, Y. SUACDNet: Attentional change detection network based on siamese U-shaped structure. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102597. [CrossRef]
45. Wang, Y.; Hong, D.; Sha, J.; Gao, L.; Liu, L.; Zhang, Y.; Rong, X. Spectral–spatial–temporal transformers for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
46. Zhang, W.; Zhang, Q.; Ning, H.; Lu, X. Cascaded attention-induced difference representation learning for multispectral change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *121*, 103366. [CrossRef]
47. Wang, Q.; Zhang, X.; Chen, G.; Dai, F.; Gong, Y.; Zhu, K. Change detection based on Faster R-CNN for high-resolution remote sensing images. *Remote Sens. Lett.* **2018**, *9*, 923–932. [CrossRef]
48. Ding, Q.; Shao, Z.; Huang, X.; Altan, O. DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102591. [CrossRef]
49. Shu, Q.; Pan, J.; Zhang, Z.; Wang, M. DPCC-Net: Dual-perspective change contextual network for change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102940. [CrossRef]
50. Yin, H.; Weng, L.; Li, Y.; Xia, M.; Hu, K.; Lin, H.; Qian, M. Attention-guided siamese networks for change detection in high resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103206. [CrossRef]
51. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
52. Hu, K.; Li, J.; Lu, M.; Weng, L.; Xia, M. FedGCN: Federated Learning-Based Graph Convolutional Networks for Non-Euclidean Spatial Data. *Mathematics* **2022**, *10*, 100. [CrossRef]
53. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
54. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015): 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings—Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
55. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
56. Varghese, A.; Gubbi, J.; Ramaswamy, A.; Balamuralidhar, P. ChangeNet: A deep learning architecture for visual change detection. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
57. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1194–1206. [CrossRef]
58. Qian, J.; Xia, M.; Zhang, Y.; Liu, J.; Xu, Y. TCDNet: Trilateral Change Detection Network for Google Earth Image. *Remote Sens.* **2020**, *12*, 2669. [CrossRef]
59. Chu, S.; Li, P.; Xia, M. MFGAN: Multi feature guided aggregation network for remote sensing image. *Neural Comput. Appl.* **2022**, *34*, 10157–10173. [CrossRef]
60. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
61. Li, Z.; Tang, C.; Wang, L.; Zomaya, A.Y. Remote sensing change detection via temporal feature interaction and guided refinement. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
62. Zhang, X.; Yang, P.; Zhou, M. Multireceiver SAS imagery with generalized PCA. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3286180. [CrossRef]
63. Jiang, N.; Du, H.; Ge, S.; Zhu, J.; Feng, D.; Wang, J.; Huang, X. High-Resolution Azimuth Missing Data SAR Imaging Based on Sparse Representation Autofocusing. *Remote Sens.* **2023**, *15*, 3425. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.