remote sensing

MDPI

Article

MFINet: Multi-Scale Feature Interaction Network for Change Detection of High-Resolution Remote Sensing Images

Wuxu Ren [1], Zhongchen Wang [1], Min Xia [1,*] and Haifeng Lin [2]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202183370005@nuist.edu.cn (W.R.); 202212490598@nuist.edu.cn (Z.W.)

[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210000, China; haifeng.lin@njfu.edu.cn

* Correspondence: xiamin@nuist.edu.cn

Abstract: Change detection is widely used in the field of building monitoring. In recent years, the progress of remote sensing image technology has provided high-resolution data. However, unlike other tasks, change detection focuses on the difference between dual-input images, so the interaction between bi-temporal features is crucial. However, the existing methods have not fully tapped the potential of multi-scale bi-temporal features to interact layer by layer. Therefore, this paper proposes a multi-scale feature interaction network (MFINet). The network realizes the information interaction of multi-temporal images by inserting a bi-temporal feature interaction layer (BFIL) between backbone networks at the same level, guides the attention to focus on the difference region, and suppresses the interference. At the same time, a double temporal feature fusion layer (BFFL) is used at the end of the coding layer to extract subtle difference features. By introducing the transformer decoding layer and improving the recovery effect of the feature size, the ability of the network to accurately capture the details and contour information of the building is further improved. The F1 of our model on the public dataset LEVIR-CD reaches 90.12%, which shows better accuracy and generalization performance than many state-of-the-art change detection models.

Keywords: remote sensing images; change detection; transformer; self-attention mechanism; CNN

check for updates

Citation: Ren, Z.; Xia, M.; Lin, H. MFINet: Multi-Scale Feature Interaction Network for Change Detection of High-Resolution Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1269. https://doi.org/10.3390/rs16071269

Academic Editor: Mohammad Awrangjeb

Received: 26 February 2024
Revised: 28 March 2024
Accepted: 2 April 2024
Published: 4 April 2024

1. Introduction

With the development of earth observation technology and geographic information technology, remote sensing images have become more and more abundant and diverse. The widespread use of satellites, aircraft, and other sensors enables us to capture information on the Earth's surface, including features of terrain, land cover, vegetation, buildings, and other geographical objects [1]. This remote sensing technology can also obtain data in different spectral ranges, including infrared and ultraviolet spectra, which helps us understand surface features more comprehensively [2–4].

With the development of remote sensing technology and the acceleration of urbanization, the problem of change detection has become more complex. It has become an urgent challenge to detect change areas quickly and accurately from the massive amount of land cover remote sensing image data [5]. In this context, the research of building change detection technology has become crucial, as shown in Figure 1. Its main goal is to accurately identify and locate regions where semantic changes have occurred from a pair of time series remote sensing images, that is, the true change region, and suppress the influence of the pseudo-change region [6]. This technology has broad application prospects in many fields, including environmental monitoring [7], climate research [8], disaster assessment [9], agricultural management [10], urban planning [11], and water resource management.

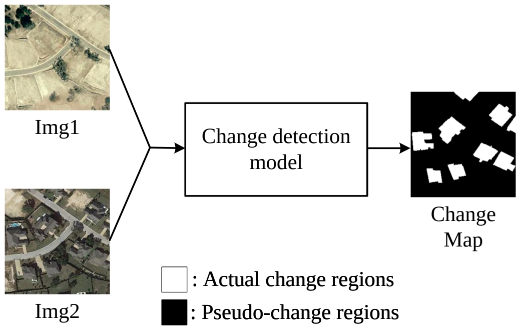

Figure 1. Graphical illustration of the change detection task.

Change detection methods in remote sensing imagery can be divided into pixel-level [12], feature-level [13], and object-level [14] methods according to the granularity of the change object. By sorting out the development process of remote sensing image change detection, the development route of international remote sensing image change detection technology can be divided into four different stages.

In the first stage, remote sensing technology faced constraints imposed by satellite and optical instrument limitations, resulting in low data quality. The hallmark of this period was the adoption of straightforward algebraic calculations or direct pixel comparisons to generate change detection results. For example, principal component analysis (PCA) [15] was extensively employed. Zhong et al. [16] proposed an unsupervised change detection method utilizing PCA and k-means clustering. This approach involved segmenting differential images into non-overlapping blocks, projecting pixels into the feature vector space, and employing k-means clustering for detection. Another notable example from this stage is change vector analysis (CVA). Liu et al. [17] introduced a novel multi-scale morphological compressed change vector analysis method. This method expanded on the spectral-based compressed change vector analysis approach by jointly analyzing spectral–spatial change information. It utilized morphological analysis to construct reconstructed spectral change vector features, preserving more geometric details.

In the second stage, machine learning methods such as support vector machines (SVMs) [18] and decision trees [19] were introduced. Volpi et al. [20] conducted research using histogram statistics as fundamental detection features, followed by the application of SVMs for land-use change detection. Im et al. [19] combined image neighborhood correlation analysis with change detection methods based on decision tree classification. Change detection methods based on machine learning algorithms have the capability to automatically extract features from large-scale remote sensing data, exhibiting excellent sensitivity to complex and subtle changes. However, these methods generally face the problem of high computational overhead.

In the third stage, object-level change detection emerged as a departure from pixel-level change detection. Unlike focusing solely on changes in individual pixels, object-level change detection emphasizes detecting changes at the level of target objects or entities. In the work by Wang et al. [21] presented a change detection approach based on objects, which integrates spectral, shape, and texture features, employing multiple supervised classifiers. The accuracy of change detection in urban environments was improved through the utilization of a weighted voting ensemble strategy. Tan et al. [22] introduced an object-based multi-feature change detection method, which uses multiple features and random forests to select features. Object-level methods usually include steps such as object extraction, feature representation, matching, and context modeling to obtain more accurate change information. However, these methods can only extract low-level features in images, which are obviously affected by factors such as radiation differences.

In the fourth stage, recent years have witnessed significant advancements in computer vision technology, with deep learning providing promising solutions to change detection problems. Traditional methods for change detection often rely on manually designed features and rules. Faced with the ever-growing volume of high-resolution remote sensing data, the performance of these methods gradually becomes limited. Deep learning techniques, particularly the application of convolutional neural network (CNN) [23] and transformer [24] models, have injected new vitality into change detection [25]. The prominence of deep learning methods in the field of change detection arises from their ability to learn features from data without the need for manual feature extraction, thereby enhancing adaptability to change patterns [26]. Through deep learning, the model can automatically capture the contextual information, textural features, and semantic information in the image.

Existing deep learning-based change detection methods lack interactive expression between bi-temporal images during the encoding phase, resulting in the isolation of bi-temporal information and the limited discernibility of actual change regions. Furthermore, in the decoding phase, the use of excessively high sampling rates and the absence of skip connections with the encoding module prevent effective multi-scale information fusion. This lack of fusion, along with poor communication of contextual information, hinders the layer-wise restoration of image features. Consequently, this leads to numerous false positives and negatives at segmented edges in the detected images [27]. Our proposed method aims to enhance the bi-temporal interaction during the feature extraction phase of Siamese models. It combines the advantages of local feature extraction from CNNs and the global feature extraction capabilities of transformers. We optimize the overall information recovery capability during the model's upsampling process to achieve high-precision, high-generalization change detection. The main contributions of our work are as follows:

1. A remote sensing image change detection network based on a multi-scale feature interaction structure named MFINet is proposed to solve the problem of insufficient target attention caused by insufficient bi-temporal interaction in change detection tasks. In the overall structure, we use a combination of a CNN encoder and a transformer decoder to make full use of the CNN's local perception and the transformer's global receptive field to effectively understand different levels of multi-source information.
2. A bi-temporal feature interaction layer (BFIL) is proposed to act as a medium for multi-level feature interaction, enhance the semantic information exchange between the same-level features of the Siamese network, and enhance the multi-temporal information communication at different time nodes. It is conducive to the model to discover the actual change regions and suppress the interference of the pseudo-change region.
3. In order to strengthen the model's perception of the fine-grained difference between the bi-temporal deep processing features, we propose the bi-temporal feature fusion layer (BFFL), which integrates rich bi-temporal deep features before image size restoration by constructing bi-temporal homologous global guidance features.

2. Related Work

2.1. CNN-Based Change Detection Methods

CNNs are favored because of their inductive bias and generalization. Zhan et al. [28] first introduced a CNN into SiameseNet as a solution for change detection. The twin network reuses the same codec structure for two temporal images, learns the bi-temporal image features in an equal way, and obtains the change information. Daudt et al. [29] introduced a twin fully convolutional network (FCN) into the end-to-end remote sensing image change detection task, and proposed three different network architectures. FC-EF uses the method of splicing dual-phase images as input, while FC-Siam-conc and FC-Siam-diff use a twin FCN structure. Peng et al. [30], based on the UNet++ encoder–decoder structure, used global and fine information to generate feature maps with high spatial accuracy. Then, the fusion strategy of multiple auxiliary outputs was used to combine the change maps of different semantic levels to generate the final change map with high accuracy. In summary, many researchers have directly transplanted classical models in semantic segmentation, such as UNet and FCN, to the field of change detection. However, the change detection task is bi-temporal, which is different from the single temporality of semantic segmentation. These models often form a twin structure by copying the existing codec structure, which lacks bi-temporal interaction and is difficult to adapt to change detection datasets with large time spans. Therefore, Zhang et al. [31] proposed IFNet, which adopts a two-stream architecture to interact with information twice, and then uses a deep supervised difference discriminant network (DDN) for change detection. In order to improve the integrity of the output change map and the internal compactness of the object, IFNet fuses the multi-level deep features of the original image with the image difference features through the attention mechanism. Yin et al. [32] proposed SAGNet, which interspersed the bi-temporal interaction scheme between the coding levels. Through the hybrid layer and the backbone network combined with the bi-temporal contextual information, the bi-temporal feature distribution is more similar, and the automatic domain adaptation between the two time domains is realized to a certain extent. Although the above methods are all based on CNNs, they mainly focus on the local perception of convolution kernels, and it is difficult to effectively model remote contextual information in bi-temporal images, which greatly limits their performance.

2.2. Transformer-Based Change Detection Methods

The research on traditional change detection tasks mainly focuses on extracting spatio-temporal contextual information by increasing the receptive field of the model, such as using dilated convolution instead of traditional convolution. However, although this method can expand the receptive field, it is usually accompanied by a huge amount of parameters and cannot really map the global features of the image. In response to these problems, models based on the self-attention mechanism have begun to emerge. Chen et al. [33] proposed a network, STANet, that emphasizes the interaction of spatio-temporal features, which closely combines the temporal information and spatial information in bi-temporal remote sensing images to capture image changes more accurately. Zhang et al. [34] used the pure swin transformer blocks to form a codec structure, and used reverse patch merging to achieve upsampling. This purely self-attention-based method shows high performance in large-scale datasets, especially for remote sensing images with high resolution and complex scenes. However, due to the lack of inductive bias, the effect is not good when training small datasets. Chen et al. [35] proposed BIT based on a CNN and transformer codec. The network only uses convolutional networks in the early stage of feature extraction, and uses the transformer module shown in Figure 2 to model the context in the compact label space in the middle and later stages. It shows efficient and effective performance in the tasks, and has obvious advantages in computational cost and model parameters.

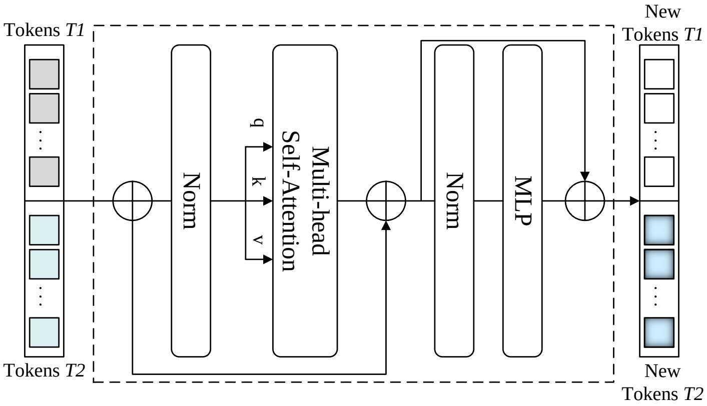

Figure 2. Illustration of transformer block.

3. Methodology

3.1. Overall Structure

The overall structure of the MFINet is shown in Figure 3. The network mainly includes two stages. The first stage is the encoding stage responsible for feature extraction. There are three members, including the multi-scale encoding layer of the backbone network ResNet18, the bi-temporal feature interaction layer, and the bi-temporal feature fusion layer. The main function of the bi-temporal feature interaction layer is to receive the output of each layer of the twin ResNet18. These outputs are features extracted from remote sensing images taken at two different time points, including changes in the target area and background information. The bi-temporal feature interaction layer allows the network to periodically focus on pixels at different time points and assign weights according to their importance. This helps one identify and capture the change area and the correlation between the bi-temporal images. The structure of the bi-temporal feature fusion layer is dual-input single-output, receiving the deepest information from the twin ResNet18, helping the network to refine the underlying features of low-resolution high channels, so as to explore the channel information that is beneficial to distinguish the change area. The second stage is the decoding stage responsible for feature size recovery. There are two components, including the transformer decoding layer and the classifier before output. In order to combine shallow detail information and deep semantic information, the difference feature maps of the two-way encoding blocks are given to the corresponding decoding blocks by skip connection. In addition, a classifier is used as a post-processing module at the end of the decoding layer to achieve binary classification.

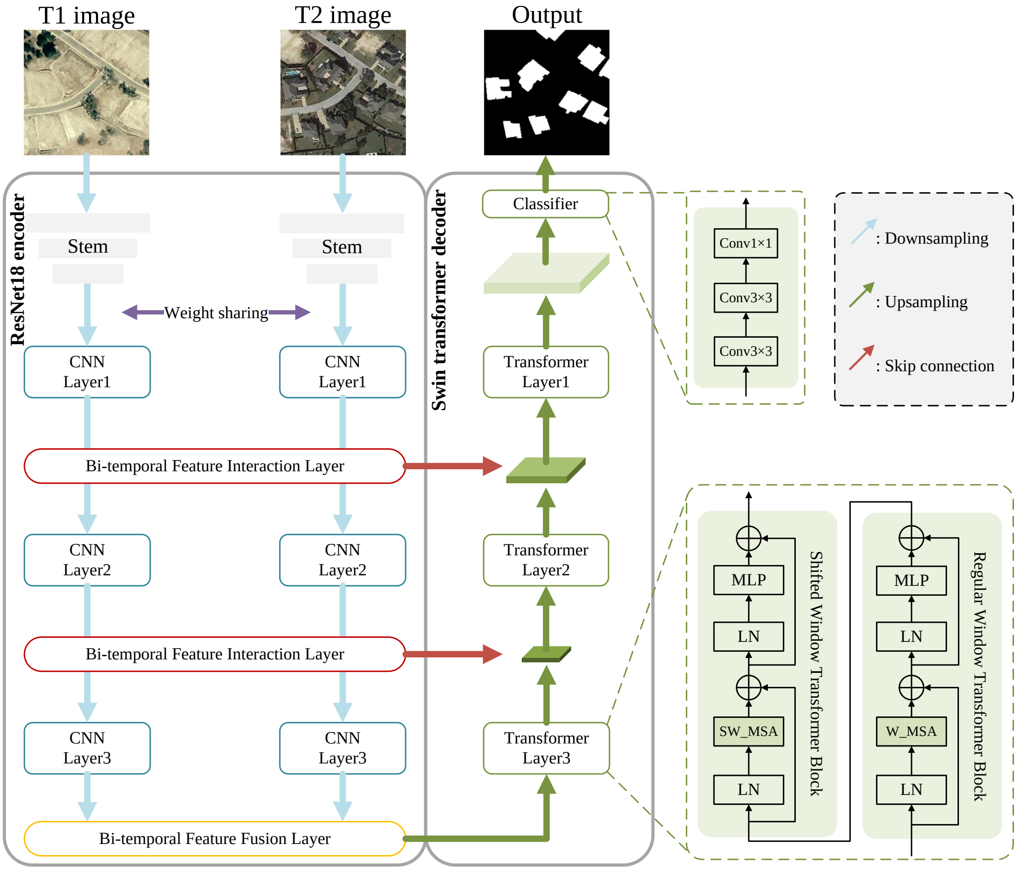

Figure 3. The overall structure diagram of the multi-scale feature interaction network; the internal structure of the transformer layer and the classifier are displayed in the green dotted box.

3.2. Bi-Temporal Feature Interaction Layer

Because there are many pseudo-changes in remote sensing images taken at different times, such as care differences and vegetation color differences caused by seasonal changes [36], in this study, as shown in Figure 4, we introduce an important module, the bi-temporal feature interaction layer (BFIL), to meet those challenges. The core task of this module is to allow the deep learning network to effectively communicate the features of images at different times, especially in the case of unbalanced actual change samples and pseudo-change samples, to suppress the interference of task-independent information [37], and to associate similar regions in different time periods, thereby improving the coding accuracy of each scale in the model encoding stage. The module consists of a pair of symmetrical transformer blocks based on self-attention, and the interaction mode is the exchange of the self-attention sequence.

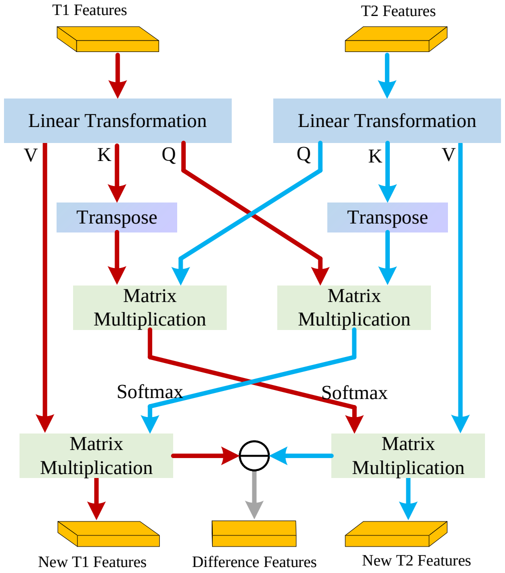

Figure 4. Structure diagram of the bi-temporal feature interaction layer.

Specifically, if we set the input single-temporal feature $f_n \in \mathbb{R}^{C\times H\times W}$ from the temporal n, then the feature will first be mapped to three identical linear transformer layers. In the layer, the original pixel matrix will first compress the channel to fuse the multi-channel features, and then expand into a self-attention vector. According to the functions that will be assigned in the future, these three generated sequences are called query vector $Q_n \in \mathbb{R}^{\frac{C}{2}\times L}$, key vector $K_n \in \mathbb{R}^{\frac{C}{2}\times L}$, and value vector $V_n \in \mathbb{R}^{\frac{C}{2}\times L}$. The process of generating sequences can be expressed by the following formula:

$$Q_n, K_n, V_n = Reshape(Linear(f_n)), \tag{1}$$

where C represents the number of channels in the feature map and three vectors. H represents the height of the feature map. W represents the width of the feature map. $L = H \times W$ represents the generated vector sequence. $Linear(\cdot)$ represents the linear layer used to change the channel. $Reshape(\cdot)$ represents the operation of the matrix changing into a vector sequence. After obtaining the triple vector sequence, the key vector performs matrix multiplication with the query vector after transpose, which can calculate the similarity score between each query vector and key vector, and convert the score through the softmax activation function to the weight A_n, which is used to weight the calculated value vector. This way of assigning attention weight to yourself can be expressed by the following dot product formula:

$$A_n = softmax({K_n}^T Q_n), \tag{2}$$

$$f_n{}' = V_n A_n. \tag{3}$$

Taking $n = 1$ and $n = 2$ as examples, the structure of the bi-temporal feature interaction layer is introduced. In this layer, we allow information to interact and pass between two tenses to better understand image changes.

First, we generate separate query vectors for n = 1 and n = 2, respectively. These query vectors represent specific information at different time points. Then, we exchange the query vectors and apply the query vectors of n = 1 to n = 2, and vice versa. In this way, we can realize the information interaction between two tenses. Next, we use the exchanged query vector together with the corresponding temporal key vector to calculate the similarity score of the elements between different temporals. These similarity scores are used to determine the correlation of different elements between the two tenses for information transmission. Finally, we use these similarity scores as the weights of the self-attention mechanism, and exchange them again to weight the value vectors of each other's tenses. This produces two outputs, ${f_1}'$ and ${f_2}'$, that fuse multi-temporal information. The following two sets of dot product formulas can express the above interaction process:

$$A_1 = soft\max({K_1}^T Q_2) \tag{4}$$

$$A_2 = soft\max({K_2}^T Q_1) \tag{5}$$

$${f_1}' = V_2 A_1 \tag{6}$$

$${f_2}' = V_1 A_2 \tag{7}$$

The existing feature interaction methods often directly perform bi-temporal interaction at the feature level. For example, the FC-CD series methods [29] interact with features through pixel-level subtraction and channel cascade. This method easily leads to semantic information confusion, making it difficult for the model to distinguish the similarities and differences between the two groups of pictures. By exchanging attention-related queries and key value information between two temporals, the BFIL can bridge the feature information of another branch while retaining the single temporal feature. The self-association and the guidance of parallel branches enhance the global attention of the model across the time domain to a certain extent and suppress the interference of pseudo-changes.

3.3. Bi-Temporal Feature Fusion Layer

In the field of change detection, common bi-temporal feature fusion methods include difference maps and transformation vectors, which aim to capture the change information between two moments. However, they mainly use simple non-parametric operations, such as pixel-level subtraction, pixel-level addition, channel concatenation, or bilinear pooling, resulting in a low matching degree of multi-channel information in bi-temporal scenarios, which is easy to confuse with feature information. Considering the low resolution and high channel properties of deep features, we propose the bi-temporal feature fusion layer (BFFL) in Figure 5. This layer can extract the global attention weight to summarize the bi-temporal features. The generation of the global attention matrix A_g can be expressed as

$$A_g = \sigma(Conv_{1\times1}(GELU(AvgPool(Concat[f_1, f_2])))), \tag{8}$$

where f_1 and f_2 represent the input bi-temporal features. $Concat[\cdot]$ represents the channel cascade. $AvgPool(\cdot)$ represents average pooling. $GELU(\cdot)$ represents the GELU activation function [38]. $Conv_{1\times1}$ represents 1×1 convolution operation. $\sigma(\cdot)$ represents the Sigmoid activation function. These weights are used to adjust the existing dual-branch original features. By performing pixel subtraction on the attention matrix corresponding to the bi-temporal and the original features after the compressed channel, it is helpful to deeply mine the potential difference features. Finally, the obtained bi-temporal features are integrated through channel cascades to improve the information richness of bi-temporal features. The formula of the feature fusion operation is expressed as follows:

$$f_{out} = Concat\left[\left|Conv_{1\times1}(A_g f_1) - f_2\right|, \left|Conv_{1\times1}(A_g f_2) - f_1\right|\right]. \tag{9}$$

This layer combines global attention and simple difference operations so that our fusion layer can capture the subtle differences of the transformation more carefully and comprehensively, thereby improving the accuracy of change detection. Existing fusion feature algorithms, such as bilateral guided aggregation layer [39] and ensemble channel attention module, rely on high-channel fusion of multi-scale features, resulting in huge computational overhead. BFFL is more flexible and efficient, and can better adapt to transformations in complex scenes. Through the targeted operation of deep features, our method shows stronger discrimination when dealing with high-channel deep features, thus providing a more powerful feature expression for change detection tasks.

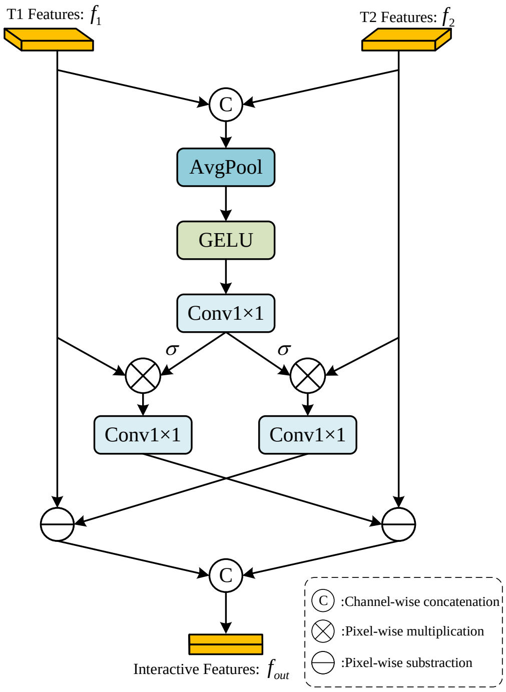

Figure 5. Structure diagram of bi-temporal feature fusion layer.

3.4. Multi-Scale Decoding Layer Based on Transformer

Change detection involves capturing the changes in images at different times, and these changes usually contain rich details and semantic information. The encoding module maps the input image to a high-level abstract feature space, but simple deconvolution or bilinear interpolation may lead to reduced resolution and loss of information. Therefore, in order to restore high-level abstract features to the original input space, a decoder with a global receptive field is a wise choice [27].

Compared with traditional CNNs, using a swin transformer block based on self-attention mechanisms as a decoding module has unique advantages. In Figure 3, the swin transformer block introduces the concept of a mobile attention window, which enables the network to capture long-distance dependencies in the global range with a small computational cost. This solves the problem that the computational complexity of ViT increases with the square of the image size, which is conducive to processing large-scale images better. Compared with CNNs, the swin transformer provides better scalability and global modeling capabilities while maintaining efficient performance [34]. This feature is particularly important for change detection tasks, because the impact of changes usually involves a wide area of the image. The mathematical expression of the continuous swin transformer block using the shift window division method is

$$\hat{y}_l = W\_MSA(Linear(y_{l-1})) + y_{l-1}, \tag{10}$$

$$y_l = MLP(Linear(\hat{y}_l)) + \hat{y}_l, \tag{11}$$

$$\hat{y}_{l+1} = SW\_MSA(Linear(y_{l+1})) + y_l, \tag{12}$$

$$y_{l+1} = MLP(Linear(\hat{y}_{l+1})) + \hat{y}_{l+1} \tag{13}$$

where $\hat{y}_l$ and y_l represent the output characteristics of the $(S)W\_MSA$ module and the multi-layer perceptron (MLP), respectively. $W\_MSA(\cdot)$ and $SW\_MSA(\cdot)$ represent multi-head self-attention based on conventional window partition and multi-head self-attention based on moving window partition, respectively. $Linear(\cdot)$ represents the linear layer.

Bilinear interpolation is used to recover the features layer by layer between the coding modules. In addition, as shown in Figure 3, skip connection is used to associate the encoding layer with the decoding layer. Specifically, the output of each decoding block is reduced to a feature matrix and added with the difference output of the bi-temporal feature interaction layer, so as to realize the weighting of the local difference feature to the global feature and further reduce the loss of difference information in the upsampling process.

4. Experiment

4.1. Datasets

4.1.1. LEVIR-CD

As shown in Figure 6, the dataset uses large-scale and high-resolution remote sensing images obtained by Google Earth, and the target changes include various types of buildings in urban and rural areas such as homes and warehouses. Containing multiple sets of image data, the time span between different groups varies, and the introduction of seasonal changes and changes caused by illumination can effectively verify the network's ability to focus on target changes. The details of the dataset are shown in Table 1.

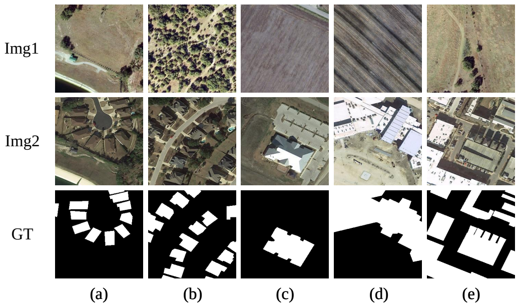

Figure 6. LEVIR-CD diagram. Each column of (**a**–**e**) represents a sample. The first and second rows show the bi-temporal remote sensing images, and the third row shows the ground truth.

4.1.2. GZ-CD

As shown in Figure 7, the dataset captures Guangzhou in 2006 and 2019 using 19 pairs of remote sensing images obtained from Google Earth. The target changes in the dataset include various types of buildings. It is worth noting that GZ-CD contains a small number of samples, so the degree to which the network relies on a large number of labeled data can be checked by comparing the level with other datasets [40]. The details of the dataset are shown in Table 1.

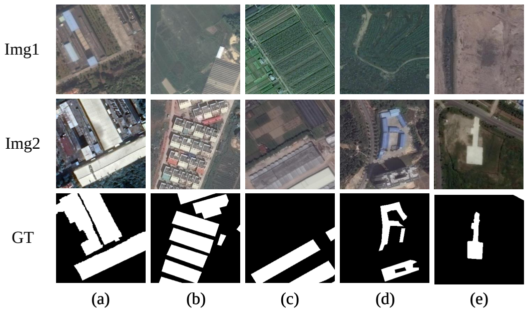

Figure 7. GZ-CD diagram. Each column of (**a**–**e**) represents a sample. The first and second rows show the bi-temporal remote sensing images, and the third row shows the ground truth.

4.1.3. Lebedev Dataset

As shown in Figure 8, the dataset was collected from Google Earth, and the shooting objects included multiple sets of remote sensing images from the same geographical area but different seasons, and the shooting resolution was inconsistent. Actual change regions included man-made objects such as roads, cars, buildings, and natural objects such as individual trees and forests. Significant seasonal differences led to significant brightness changes, which made it difficult for the network to distinguish between target changes and background changes [41]. The details of the dataset are shown in Table 1.

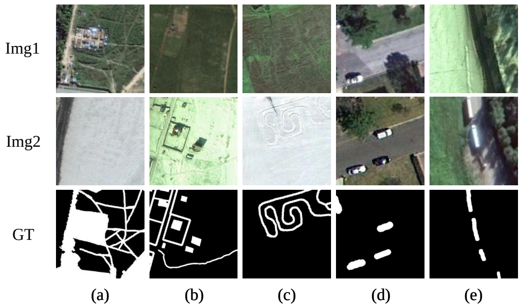

Figure 8. Lebedev dataset diagram. Each column of (**a**–**e**) represents a sample. The first and second rows show the bi-temporal remote sensing images, and the third row shows the ground truth.

Table 1. The main parameter information of the three datasets we used.

Dataset	Size	Resolution	Number of Pixels			Number of Images		
	(pixel)	(m/pixel)	Actual Change	Pseudo-Change	Ratio	Train	Validation	Test
LEVIR-CD	256 × 256	0.5	30,913,975	637,028,937	1:20.61	7120	1024	2048
GZ-CD	256 × 256	0.55	20,045,119	200,155,821	1:10.01	2504	313	313
Lebedev	256 × 256	0.03–2	134,068,750	914,376,178	1:6.83	10,000	3000	3000

4.2. Implementation Details

In terms of hardware, our experiments were configured by Intel Core i5-13600 CPU and NVIDIA RTX3080 GPU. In terms of software, Python (3.9) and Pytorch (1.10) were used. We used Binary cross entropy (BCE) loss, which is a commonly used loss function in binary change detection tasks. It combines the Sigmoid activation function with BCE loss to make the calculation more stable and efficient. The optimizer used Adam. During the network training, we used the ploy method to dynamically change the learning rate. The initial learning rate (lr_{base}) was set to 0.001. Since most of the networks converge to the minimum loss at about 200 iterations, the maximum training iteration (max_epoch) was 250, and the batch size was set to 16. The learning rate of each epoch was calculated as follows:

$$lr_{base} \times (1 - \frac{epoch}{max\_epoch}) \tag{14}$$

Five typical indicators were used to evaluate the performance of change detection, and the higher the value, the better. Four of them were used to evaluate target changes: Precision (P), Recall (R), Intersection over Union (IoU), and $F1$ score; two indicators were used to evaluate the overall classification accuracy: Overall Accuracy (OA). Formally, the five indicators are defined as

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = \frac{2}{P^{-1} + R^{-1}} \tag{17}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{18}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

where TP, TN, FP, and FN represent the quantities of true positives, true negatives, false positives, and false negatives, respectively.

4.3. Ablation Experiments on LEVIR-CD

In this section, we perform ablation research on LEVIR-CD and GZ-CD, add and subtract modules on the basis of the backbone network, and perform experiments one by one to evaluate the effectiveness of using each module in the encoding and decoding stages. Table 2 shows the results of the ablation test. The training parameters of all models are exactly the same.

Table 2. Ablation experiment on LEVIR-CD.

Method	LEVIR-CD		GZ-CD	
	F1 (%)	**IoU (%)**	**F1 (%)**	**IoU (%)**
Backbone	86.54	78.41	82.70	71.45
Backbone + BFIL	87.53	80.95	84.09	73.97
Backbone + BFIL + BFFL	88.11	81.93	84.90	74.19
Backbone + BFIL + BFFL + Dec. (CNN)	89.96	82.12	85.59	74.44
Backbone + BFIL + BFFL + Dec. (Transformer)	90.12	82.33	86.08	74.87

1. The influence of BFIL: It is difficult for a simple twin CNN network to discover the common and different features of bi-temporal features, and the ability of bi-temporal mutual understanding will become worse as the number of layers deepens. Therefore, we added a BFIL to the backbone network to strengthen the interactive attributes of bi-temporal features, and used the attention weight as an interactive means. The experimental results show that the BFIL can help the network to improve the accurate detection of changing targets in the coding stage. For LEVIR-CD, F1 increased by 0.99% and IoU increased by 2.54%. For GZ-CD, F1 increased by 1.39% and IoU increased by 2.52%.
2. The influence of BFFL: The fusion operation of deep bi-temporal features is a great test of the lightweight degree and differential feature extraction ability of the module. It is easy to confuse features using simple pixel subtraction or channel cascade, while BFFL reduces the occurrence of feature confusion through multiple residual connections. The experimental results show that the BFFL bi-temporal feature fusion significantly increases the segmentation accuracy of the changed region features. For LEVIR-CD, F1 increased by 0.58% and IoU increased by 0.98%. For GZ-CD, F1 increased by 0.81% and IoU increased by 0.22%.
3. The influence of decoder selection: We compared two kinds of decoder methods. One is ResNet18, which is consistent with the encoder, and the other is the swin transformer used in our model. In terms of experimental results, the improvement in indicators in the changing region is limited. The F1 for LEVIR-CD increased by 0.16%, and IoU increased by 0.21%. For GZ-CD, F1 increased by 0.49% and IoU increased by 0.43%.

4.4. Comparative Experiments on Different Datasets

We evaluated the multifaceted performance of MFINet by comparing it with eleven competitive change detection methods on two datasets. The methods involved in the comparison included the classic FC-CD series based on twin fully convolutional neural networks and some mainstream change detection models combined with multi-class visual algorithms in recent years. The types can be divided into two categories. Firstly, there are models based on CNNs and traditional attention mechanisms. For example, the FC-CD series includes FC-EF, FC-Siam-diff, and FC-Siam-conc. Based on the improved decoder–

encoder structure of Unet++_MSOF and SNUNet [42], IFNet uses channel attention and spatial attention to optimize the feature weight distribution in the process of multi-scale skip connections. SAGNet and SAFNet [43] add a bi-temporal interaction layer between the encoding layers to communicate the semantic information of the twin branches. Secondly, there are models combining transformers and self-attention mechanisms, such as STANet, which models spatio-temporal relationships through multi-scale pooling and self-attention mechanisms. DASNet [44] introduces a dual attention mechanism to capture long-distance dependencies and enhance feature representation to improve the recognition performance of the model. BIT uses a CNN in the initial feature extraction, and uses a transformer encoder and decoder to correlate bi-temporal information in the form of sequences in the middle and late stages. These methods have achieved competitive performance on various change detection datasets. Figures 9–11 qualitatively show the prediction graphs of each method on three datasets, where different colors are assigned to identify the correctness or inaccuracy of the detection, including TP (white), TN (black), FP (red), and FN (green).

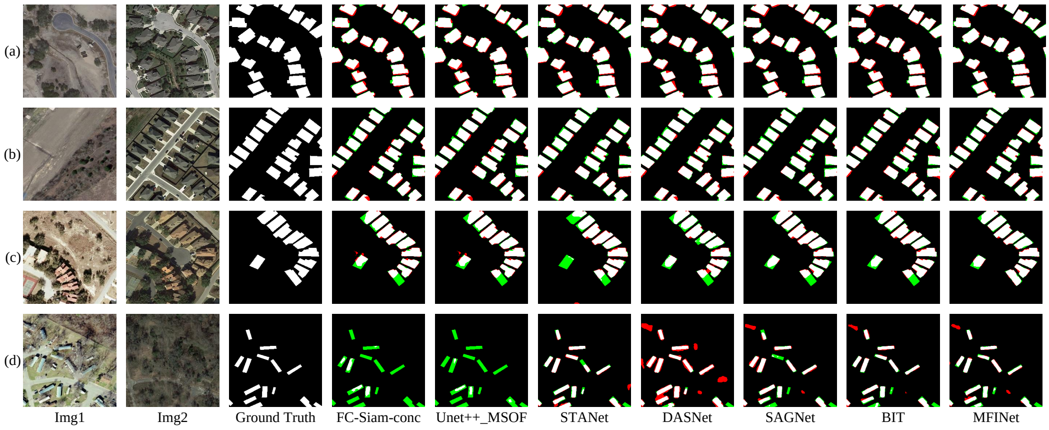

Figure 9. The quantitative performance visualization of different methods on LEVIR-CD. (**a**–**d**) denote the prediction results of all comparison methods for different samples. In the color classification, the true positive is white, the true negative is black, the false positive is red, and the false negative is green.

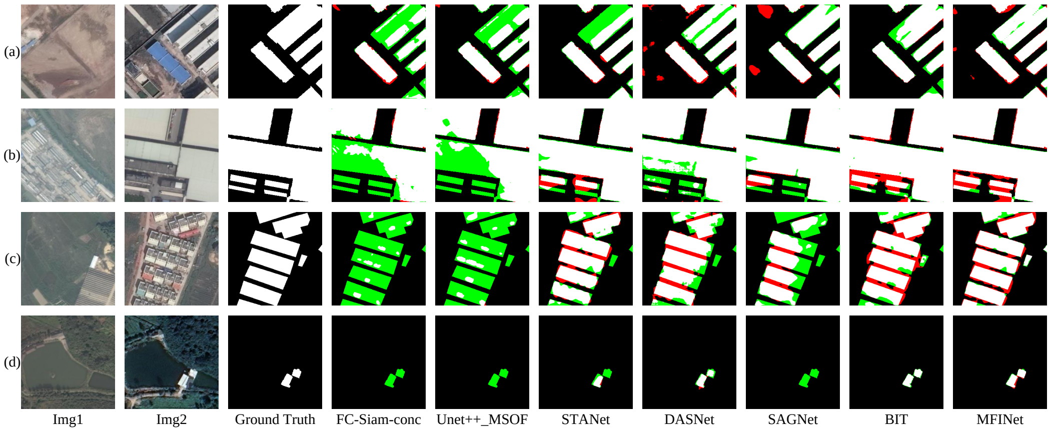

Figure 10. The quantitative performance visualization of different methods on GZ-CD. (**a**–**d**) denote the prediction results of all comparison methods for different samples. In the color classification, the true positive is white, the true negative is black, the false positive is red, and the false negative is green.

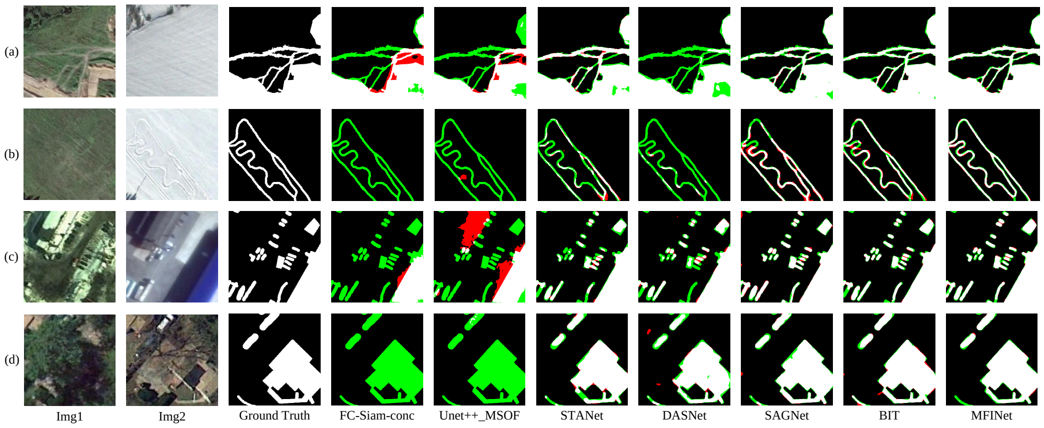

Figure 11. The quantitative performance visualization of different methods on the Lebedev dataset. (**a**–**d**) denote the prediction results of all comparison methods for different samples. In the color classification, the true positive is white, the true negative is black, the false positive is red, and the false negative is green.

4.4.1. Comparative Experiments on LEVIR-CD

Table 3 shows the performance of different models on multiple performance indicators in detail. MFINet achieves the best performance in Recall, F1-score, IoU, and OA. However, the BIT using the transformer codec is slightly higher in precision than our proposed model. This shows that MFINet is ahead of other models in comprehensive performance. In transformer-based models, such as BIT, STANet, and DASNet, although Precision and Recall are slightly lower than those in CNN-based models such as SNUNet, SAGNet, and SAFNet, OA remains at a very high level. This shows that the transformer-based model has advantages in global correlation, especially when training large-scale datasets such as LEVIR-CD. However, due to the lack of a twin structure, FC-EF leads to the confusion of the input bi-temporal features, thus showing low accuracy. In contrast, FC-Siam-conc, FC-Siam-diff, and Unet++_MSOF adopt a simple bi-temporal interaction and lack a multi-scale attention mechanism, so the model does not pay enough attention to changing targets. SAGNet and SAFNet are modeled in a multi-scale interactive manner. However, due to the lack of a global feature extraction module with a large receptive field, it is difficult to establish long-distance semantic associations when processing high-resolution remote sensing images. In summary, MFINet performs well in both comprehensive performance and processing ability of high-resolution images, showing its superiority in remote sensing image change detection tasks.

Table 3. The comparison results of different comparison models in the LEVIR-CD test set (bold numbers represent the optimal results).

Method	P (%)	R (%)	F1 (%)	IoU (%)	OA (%)
FC-EF	86.91	80.17	83.42	72.01	97.29
FC-Siam-diff	89.56	83.41	86.31	75.99	98.67
FC-Siam-conc	88.17	84.64	86.37	76.01	98.77
Unet++_MSOF	89.47	85.37	87.19	78.10	98.51
IFNet	89.74	85.26	87.34	78.23	98.70
STANet	90.53	84.68	87.51	77.79	98.22
DASNet	90.91	87.70	88.48	80.02	98.99
SNUNet	90.89	88.31	89.28	80.55	98.93
BIT	**92.67**	87.61	89.32	80.72	99.00
SAGNet	91.33	86.95	88.65	81.59	98.72
SAFNet	91.60	88.70	89.43	81.66	98.95
MFINet (Ours)	92.09	**89.02**	**90.12**	**82.33**	**99.21**

In Figure 9, we select some individuals with representative structures and ideas in the comparison models, and present four sets of detailed visual comparison maps. These illustrations aim to highlight the advantages of MFINet over other models in different scenarios. Firstly, (a) and (b) show houses with similar styles and no obvious vegetation interference. Although the two images were captured under different bi-temporal illumination conditions, MFINet significantly reduces the negative impact of illumination changes on the detection results by making full use of multi-scale bi-temporal feature interaction modeling. In particular, it is worth noting that MFINet has excellent recognition ability for building shadows, and the false detection rate is significantly lower than that for other comparison models. Secondly, the images in (c) involve the interference of dense vegetation. Plants are usually regarded as pseudo-changes in change detection tasks, and all contrast models miss a corner of the building because of tree cover. On the contrary, our proposed model uses transformer decoding to achieve long-distance feature association, and successfully restores the semantic information blocked by vegetation, so there is no large area of missed detection. Finally, the images in (d) involve container rooms of different sizes and irregular distribution. In this scenario, the models participating in the comparison are susceptible to shadows and are prone to miss small targets. In contrast, our proposed model guarantees a low missed detection rate and shows its strong ability to deal with complex scenes and small targets. These visualization results intuitively demonstrate the robustness and generalization of MFINet in different scenarios. Its ability to deal with illumination changes, vegetation occlusion, and small targets makes it perform well in remote sensing image change detection tasks.

4.4.2. Comparative Experiments on GZ-CD

From the data in Table 4, MFINet achieves the best performance on Precision, Recall, F1-score, IoU, and OA. This result significantly highlights the excellent performance of our proposed model in remote sensing image change detection tasks. It is worth noting that in the horizontal comparison, BIT, STANet, and DASNet perform relatively well on LEVIR-CD, while they perform much worse on small-scale datasets such as GZ-CD. This phenomenon reveals that the transformer-based model has some challenges when dealing with small-scale datasets. The number of parameters of such models is usually large, and they rely more on a large number of labeled data for support or pre-training. When the training set is small, these models may face the problem of overfitting, that is, they rely too much on the details of the training data and have difficultly generalizing unseen data. On the contrary, models based on traditional convolutional neural networks and multi-level skip connections, such as SAGNet, SNUNet and Unet++_MSOF, perform better on small datasets. One of the reasons is that these models have the characteristics of local connection and weight sharing, which makes CNNs more robust in learning features. In the context of small-scale datasets, this robustness enables traditional convolutional neural networks to capture the characteristics of the data better, thereby achieving better performance. In general, the superior performance of MFINet is not only reflected in large-scale datasets, but also in small-scale datasets. This further verifies the robustness and generalization of our proposed model, which gives it application potential in remote sensing image change detection tasks of different scales and complexities.

In Figure 10, we select some individuals with representative structures and ideas in the comparison models, and present four sets of detailed visual comparison maps. These illustrations aim to highlight the advantages of MFINet over other models in different scenarios. Firstly, for Figure 10a,b, the land cover changes significantly due to the long shooting time interval. Although the actual change region is relatively easy to identify, there are also many pseudo-change regions that are easily misjudged as actual changes. Unusual factors such as hardened land color or shadows lead to high false detection rates in many models. This further highlights the challenges of the model in dealing with complex land cover changes. In this regard, MFINet successfully reduces the false detection rate through effective feature learning and bi-temporal interaction, and has a more accurate ability to distinguish between true and false changes. Secondly, for Figure 10c, both tenses contain buildings, but the small buildings in Img1 are removed in Img2. This bi-temporal image is a typical case to test the bi-temporal interaction ability of the model. All the comparison models visually missed the small, white building on the right side, and MFINet successfully achieved accurate detection. This shows that MFINet has advantages in capturing small details in the spatio-temporal changes of images, which helps to better understand and utilize temporal information. Finally, in Figure 10d, the target size involved is small and easily ignored. Relying on the advantages of global feature extraction, the transformer-based method can identify the approximate area of small targets, but it is also accompanied by more missed detection. In contrast, our proposed model achieves the lowest missed detection rate, further confirming the superiority of MFINet in small target detection. These visualization results provide an intuitive confirmation of our model performance, and also highlight the advantages of MFINet over other models in different scenarios and challenges.

Table 4. The comparison results of different comparison models in the GZ-CD test set (bold numbers represent the optimal results).

Method	P (%)	R (%)	F1 (%)	IoU (%)	OA (%)
FC-EF	85.16	61.62	72.33	56.95	95.26
FC-Siam-diff	84.20	58.76	69.22	56.70	95.51
FC-Siam-conc	87.43	61.82	72.64	56.98	95.36
Unet++_MSOF	87.91	72.84	81.13	73.90	97.55
IFNet	85.65	61.28	76.91	69.52	96.22
STANet	84.95	67.61	79.37	68.92	96.80
DASNet	86.71	77.97	83.23	73.08	96.93
SNUNet	87.92	83.86	85.26	74.38	97.09
BIT	87.10	72.90	84.67	73.90	96.60
SAGNet	88.00	80.66	84.01	73.32	97.34
SAFNet	87.59	83.93	84.91	73.28	97.51
MFINet (Ours)	**88.12**	**84.20**	**86.08**	**74.87**	**97.70**

4.4.3. Comparative Experiments on Lebedev Dataset

From the data in Table 5, MFINet achieves the best performance on Precision, Recall, F1-score, IoU, and OA. This result significantly highlights the excellent performance of our proposed model in remote sensing image change detection tasks. The transformer-based model performs well on large-scale Lebedev datasets but still has defects. Although the Precision of BIT is very high, it is easily affected by the imbalance of dataset samples, resulting in poor Recall. STANet's multi-scale spatio-temporal attention can effectively focus on targets of different sizes, but the low depth limits the performance. The overall accuracy of DASNet and SNUNet is high, indicating that the use of traditional channel attention and multi-scale fusion is conducive to segmenting multiple types of small targets, but the global retrieval ability is still inferior to SAGNet and MFINet, which refer to dual-temporal self-attention interactions. The FC-CD series models and the UNet-based detection model have the problem of confusing bi-temporal feature semantics. There are serious intra-class inconsistencies in the Lebedev dataset with high target diversity, and the detection performance is extremely poor.

In Figure 11, we choose a representative method in the comparison model and present four sets of detailed visual comparison diagrams. These illustrations aim to highlight the advantages of MFINet over other models in different scenarios. First, for Figure 11a,b, the detection objects include two types of roads, soil roads and snow roads. The MFINet of bi-temporal interactive modeling can distinguish pixel-level difference regions and explore shadows that are difficult for the human eye to see. For Figure 11c,d, the detection objects include vehicles and buildings, and the size of the detection objects varies greatly. In particular, vehicles at long distances are difficult to detect using the model. However, thanks to the transformer decoder's comprehensive restoration of feature details, MFINet can effectively restore target edge information and small targets.

Table 5. The comparison results of different comparison models in the Lebedev dataset test set (bold numbers represent the optimal results).

Method	P (%)	R (%)	F1 (%)	IoU (%)	OA (%)
FC-EF	89.03	61.63	70.87	55.76	94.56
FC-Siam-diff	89.98	63.53	74.47	59.32	94.86
FC-Siam-conc	89.74	60.49	72.26	56.57	94.52
Unet++_MSOF	93.84	88.6	92.57	88.12	95.99
IFNet	95.71	89.66	92.9	88.87	97.05
STANet	96.02	90.65	93.68	88.10	98.56
DASNet	96.55	92.31	94.51	89.00	98.61
SNUNet	96.32	92.42	94.33	89.27	98.69
BIT	96.76	94.28	95.74	83.74	98.03
SAGNet	96.59	95.33	95.96	92.23	99.05
SAFNet	96.25	94.80	95.92	91.96	99.02
MFINet	**96.81**	**96.44**	**96.62**	**93.40**	**99.29**

4.5. Discussion

4.5.1. Comprehensive Efficiency Analysis of the Models

This paper aims to achieve high-precision detection while reducing computational complexity. Therefore, for LEVIR-CD, we conducted a comprehensive analysis and comparison of the network from multiple perspectives, including floating-point operations (FLOPs), number of parameters (Params), inference time, and F1-score. The unit of flops is Memory Access Cost (Mac). We randomly selected 1000 images of 256 × 256 pixels in the validation set for the inference operation, and averaged all the results to evaluate the inference time of the model. The specific results are shown in Table 6. MFINet performed well on multiple performance indicators. Although the FC-CD series had a slight advantage in the F1 value, MFINet was significantly better than other models involving transformers in FLOPs and Params, achieving the highest F1 value. This shows that MFINet greatly reduces the computational burden while achieving high performance, and provides a more efficient solution for practical applications. However, it is worth noting that because our model uses GELU as the activation function many times in the bi-temporal feature fusion layer and decoder, the inference time does not show an advantage over other comparison models. Although it is competitive in computational cost, it also suggests that we can consider the choice of activation function when further optimizing the model to further improve the inference speed. In general, MFINet achieves excellent detection results with less computational cost, and is more friendly to hardware devices. This provides a more feasible choice for actual deployment, especially in resource-constrained environments. MFINet shows potential in high-performance target detection.

4.5.2. Model Characteristics and Future Prospects

In the case of remote sensing image change detection, it is a complex task to construct a robust change detection method, which requires not only the extraction of high-level semantic information to obtain the cognition of the change area, but also the acquisition of local and global intrinsic features. Our network innovatively uses a lightweight ResNet18 backbone network and the bi-temporal feature interaction layer in the coding stage, which effectively integrates multi-temporal remote sensing image information. In the decoding stage, the transformer decoding layer and the classifier further improve the effect of feature size recovery, so as to realize the fast and accurate change detection of remote sensing images.

Table 6. Comparative experiments of multiple efficiency indicators of the models.

Method	Flops (G)	Param (M)	Inference (ms/picture)	F1 (%)
FC-EF	1.19	1.35	2.29	83.42
FC-Siam-diff	2.33	1.35	9.82	86.31
FC-Siam-conc	2.33	1.55	10.41	86.37
Unet++_MSOF	18.04	7.76	18.83	87.19
IFNet	77.88	35.99	13.02	87.34
STANet	18.03	16.94	13.16	87.51
DASNet	107.69	57.36	19.27	88.48
SNUNet	43.94	12.03	12.51	89.28
BIT	25.92	11.99	14.03	89.32
SAGNet	12.25	32.23	16.37	88.65
SAFNet	14.47	40.22	18.30	89.43
MFINet (Ours)	6.89	4.95	15.62	90.12

Although our model achieved remarkable results in high-performance change detection, we should also admit that it is still highly dependent on a large number of labeled data points as support. In the future, we plan to improve the performance of the model through various optimizations, especially to improve the generalization ability on small datasets. In this regard, we will actively explore the introduction of more unsupervised learning techniques to reduce the dependence on labeled data, thereby improving the adaptability and robustness of the model. Unsupervised learning methods can help the model learn feature representations from unlabeled data, thereby improving its ability to detect changes in various environments. On the other hand, we will also continue to study the introduction of more supervised learning and self-supervised learning methods to enhance the learning ability of the model when changing patterns. Considering the particularity of different scenarios, we will deepen the adaptability of the model in complex environments. This means that we will be committed to providing more extensive and accurate change detection services. We plan to introduce more domain-specific data in future studies to ensure that our models perform well in various complex situations.

5. Conclusions

The multi-scale feature interaction network proposed in this paper provides an innovative solution for remote sensing image change detection tasks. Different from the existing model's dependence on high-depth encoding, our model achieves efficient information interaction for multi-temporal remote sensing images through lightweight encoding and bi-temporal feature interaction. At the same time, the transformer decoding layer is introduced in the decoding stage of the network architecture, which effectively improves the recovery effect of the feature size, and makes the network capture the details and contour information of the building more accurately in the output stage. The model shows high change area detection accuracy and overall image prediction accuracy on datasets of different scales, and the computational overhead is far lower than that of similar models. It shows strong generalization ability and is suitable for remote sensing images of different scenes and time scales.

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

Funding: This work was 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. Ding, L.; Xia, M.; Lin, H.; Hu, K. Multi-Level Attention Interactive Network for Cloud and Snow Detection Segmentation. *Remote Sens.* **2024**, *16*, 112. [CrossRef]
2. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical Remote Sensing Image Change Detection Based on Attention Mechanism and Image Difference. *IEEE Trans. Geosci. Remote Sns.* **2021**, *59*, 7296–7307. [CrossRef]
3. Marin, C.; Bovolo, F.; Bruzzone, L. Building Change Detection in Multitemporal Very High Resolution SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2664–2682. [CrossRef]
4. Wang, Z.; Xia, M.; Weng, L.; Hu, K.; Lin, H. Dual Encoder–Decoder Network for Land Cover Segmentation of Remote Sensing Image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2372–2385. [CrossRef]
5. Fang, S.; Li, K.; Li, Z. Changer: Feature Interaction is What You Need for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610111. [CrossRef]
6. Diakogiannis, F.; Waldner, F.; Caccetta, P. Looking for change? Roll the Dice and demand Attention. *Remote Sens.* **2021**, *13*, 3707. [CrossRef]
7. Willis, K.S. Remote sensing change detection for ecological monitoring in United States protected areas. *Biol. Conserv.* **2015**, *182*, 233–242. [CrossRef]
8. Jin, H.; He, W.; Liu, Q.; Wang, J.; Feng, G. The applicability of research on moving cut data-approximate entropy on abrupt climate change detection. *Theor. Appl. Climatol.* **2016**, *124*, 475–486. [CrossRef]
9. Qiao, H.; Wan, X.; Wan, Y.; Li, S.; Zhang, W. A novel change detection method for natural disaster detection and segmentation from video sequence. *Sensors* **2020**, *20*, 5076. [CrossRef]
10. Lunetta, R.S.; Knight, J.F.; Ediriwickrema, J.; Lyon, J.G.; Worthy, L.D. Land-cover change detection using multi-temporal MODIS NDVI data. In *Geospatial Information Handbook for Water Resources and Watershed Management*; CRC Press: Boca Raton, FL, USA, 2022; Volume II, pp. 65–88.
11. 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. [CrossRef]
12. Rokni, K.; Ahmad, A.; Solaimani, K.; Hazini, S. A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 226–234. [CrossRef]
13. Wiratama, W.; Lee, J.; Sim, D. Change detection on multi-spectral images based on feature-level U-Net. *IEEE Access* **2020**, *8*, 12279–12289. [CrossRef]
14. Xu, L.; Jing, W.; Song, H.; Chen, G. High-resolution remote sensing image change detection combined with pixel-level and object-level. *IEEE Access* **2019**, *7*, 78909–78918. [CrossRef]
15. Maćkiewicz, A.; Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342. [CrossRef]
16. 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]
17. Liu, S.; Du, Q.; Tong, X.; Samat, A.; Bruzzone, L.; Bovolo, F. Multiscale Morphological Compressed Change Vector Analysis for Unsupervised Multiple Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 4124–4137. [CrossRef]
18. Bovolo, F.; Bruzzone, L.; Marconcini, M. A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 2070–2082. [CrossRef]
19. Im, J.; Jensen, J.R. A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sens. Environ.* **2005**, *99*, 326–340. [CrossRef]
20. Volpi, M.; Tuia, D.; Bovolo, F.; Kanevski, M.; Bruzzone, L. Supervised change detection in VHR images using contextual information and support vector machines. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *20*, 77–85. [CrossRef]
21. Wang, X.; Liu, S.; Du, P.; Liang, H.; Xia, J.; Li, Y. Object-Based Change Detection in Urban Areas from High Spatial Resolution Images Based on Multiple Features and Ensemble Learning. *Remote Sens.* **2018**, *11*, 276. [CrossRef]
22. Tan, K.; Zhang, Y.; Wang, X.; Chen, Y. Object-Based Change Detection Using Multiple Classifiers and Multi-Scale Uncertainty Analysis. *Remote Sens.* **2019**, *10*, 359. [CrossRef]
23. Rawat, W.; Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput.* **2017**, *29*, 2352–2449. [CrossRef] [PubMed]
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
25. Liu, M.; Chai, Z.; Deng, H.; Liu, R. A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4297–4306. [CrossRef]
26. Ren, H.; Xia, M.; Weng, L.; Hu, K.; Lin, H. Dual-Attention-Guided Multiscale Feature Aggregation Network for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 4899–4916. [CrossRef]

27. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. In Proceedings of the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022. [CrossRef]
28. 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]
29. 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.
30. 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]
31. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
32. 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]
33. 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]
34. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure transformer network for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
35. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection With Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
36. Feng, Y.; Jiang, J.; Xu, H.; Zheng, J. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
37. Chen, C.P.; Hsieh, J.W.; Chen, P.Y.; Hsieh, Y.K.; Wang, B.S. SARAS-net: Scale and relation aware siamese network for change detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 14187–14195.
38. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv* **2016**, arXiv:1610.02136.
39. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
40. Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; Ding, H.; Huang, X. SemiCDNet: A Semisupervised Convolutional Neural Network for Change Detection in High Resolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5891–5906. [CrossRef]
41. Lebedev, M.A.; Vizilter, Y.V.; Vygolov, O.V.; Knyaz, V.A.; Rubis, A.Y. Change detection in remote sensing images using conditional adversarial networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *XLII-2*, 565–571. [CrossRef]
42. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
43. 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. [CrossRef]
44. 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.* **2021**, *14*, 1194–1206. [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.