electronics

MDPI

Article

Building Change Detection in Remote Sensing Imagery with Focal Self-Attention and Multi-Level Feature Fusion

Peiquan Shen [1], Liye Mei [2], Zhaoyi Ye [2], Ying Wang [3], Qi Zhang [2], Bo Hong [3], Xiliang Yin [3] and Wei Yang [3,*]

[1] Electronic Information School, Wuhan University, Wuhan 430072, China
[2] School of Computer Science, Hubei University of Technology, Wuhan 430068, China; 102111136@hbut.edu.cn (Q.Z.)
[3] School of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430064, China
* Correspondence: yangwei403@wsyu.edu.cn

Abstract: Accurate and intelligent building change detection greatly contributes to effective urban development, optimized resource management, and informed decision-making in domains such as urban planning, land management, and environmental monitoring. Existing methodologies face challenges in effectively integrating local and global features for accurate building change detection. To address these challenges, we propose a novel method that uses focal self-attention to process the feature vector of input images, which uses a "focusing" mechanism to guide the calculation of the self-attention mechanism. By focusing more on critical areas when processing image features in different regions, focal self-attention can better handle both local and global information, and is more flexible and adaptive than other methods, improving detection accuracy. In addition, our multi-level feature fusion module groups the features and then constructs a hierarchical residual structure to fuse the grouped features. On the LEVIR-CD and WHU-CD datasets, our proposed method achieved F1-scores of 91.62% and 89.45%, respectively. Compared with existing methods, ours performed better on building change detection tasks. Our method therefore provides a framework for solving problems related to building change detection, with some reference value and guiding significance.

Keywords: remote sensing imagery; building change detection; focal self-attention; multi-level feature fusion

Citation: Shen, P.; Mei, L.; Ye, Z.; Wang, Y.; Zhang, Q.; Hong, B.; Yin, X.; Yang, W. Building Change Detection in Remote Sensing Imagery with Focal Self-Attention and Multi-Level Feature Fusion. *Electronics* **2023**, *12*, 2796. https://doi.org/10.3390/electronics12132796

Academic Editor: Chiman Kwan

Received: 15 May 2023
Revised: 22 June 2023
Accepted: 23 June 2023
Published: 24 June 2023

1. Introduction

As remote sensing technology has developed rapidly, building change detection has become a highly concerning issue [1]. Building change detection is the identification and localization of building changes by comparing remote sensing images at different time points. Change detection in remote sensing images plays a crucial role in various applications, such as urban development monitoring, environmental analysis, and disaster assessment [2]. As one of the basic spatial structures in cities, the spatial distribution of building changes is of substantial significance for urban planning, illegal construction monitoring, and other aspects. The identification and analysis of changes between different time periods provide valuable insights for decision-making processes. Therefore, achieving effective change detection is one of the current hotspots and difficulties in the field of remote sensing images [3]. However, change detection in remote sensing images remains a challenging task due to various factors, including sensor characteristics, image registration, noise, and complex scene dynamics.

The problem of building change detection specifically focuses on detecting changes in buildings or man-made structures between different time periods. The accurate and efficient detection of building changes is essential for urban planning, land use monitoring, and disaster management. For example, detecting new construction, demolitions, or changes in building conditions can provide critical information for city planners, insurers, and emergency responders [4]. Traditional methods for building change detection often

rely on handcrafted features and heuristics, which may not fully capture the complex patterns and variations in remote sensing images. These methods often struggle with issues such as varying illumination conditions, occlusions, and geometric distortions, leading to suboptimal results. Early building change detection methods mainly used traditional image processing methods, such as template matching [5], threshold-based [6] and morphological processing [7] methods. Gong et al. [8] calculated the spectral difference between corresponding superpixels from different multi-spectral images, and trained a convolutional neural network (CNN) to learn the difference representation of the superpixels, which was used to distinguish between changed and unchanged regions. However, this method ignores contextual information and may lead to the generation of noise. Seo et al. [9] proposed a radiometric and phenological normalization image generation method based on random forest regression, which is applicable to different types of remote sensing images and provides a novel strategy for change detection. However, it is not effective in detecting non-linear changes in some scenes. Morton et al. [10] proposed a novel change detection method based on transform invariance and orthogonal linear regression. The method employs an iterative reweighting approach to estimate invariance probability, resulting in the improved accuracy of change detection. With the advancement of deep learning, building change detection methods on the basis of deep learning have also been widely applied and researched [11]. Peng et al. [12] proposed an improved semi-supervised convolutional network based on generative adversarial networks, which combined labeled and unlabeled data in training to reduce the requirement for labeled data while utilizing available information. For remote sensing images to effectively capture change information, the model employs an adaptive network architecture that integrates multiple feature extractors and attention mechanisms. Experimental results demonstrate that compared to other methods, this approach achieves better performance and higher stability. Xiao et al. [13] proposed an innovative method that builds upon object-level feature extraction and co-segmentation techniques. This method detects building changes by extracting object features and performing co-segmentation on multiple temporally sensed images, fully leveraging the multi-temporal information and avoiding the problem of missed or false detections caused by subtle differences between different time periods. Experimental results demonstrate that this method improves building change detection accuracy and has practical applications. Zhang et al. [14] proposed a feature-output space dual-alignment change detection method (FODA). This method reduces the negative impact of pseudo-changes by focusing on the correlation between constant regions in multi-temporal images. Through adversarial learning, it can enhance the ability to extract features from unchanged regions and recognize pseudo-changes. Compared with existing methods, FODA can considerably enhance change detection performance and help to effectively decrease pseudo-change issues. Zhang et al. [15] proposed an innovative model using multiscale and attention mechanisms. A multiscale attention module in this model captured multiscale semantic information and built an end-to-end dual-scale attention architecture. A dual-threshold automatic data balancing rule was also designed to alleviate severe data distribution skewness. The results of the experiments indicate that this method performs well when it comes to detecting details.

Transfer learning has emerged as a major approach for addressing the challenges of deep learning in remote sensing applications. Remote sensing tasks often suffer from limited labeled data, complex environmental conditions, and the need for specialized feature extraction [16]. In such scenarios, transfer learning offers a practical solution by leveraging knowledge learned from pretraining on large-scale datasets in related domains. Zhong et al. [17] proposed an SAR target image classification method based on transfer learning and model compression. They utilized a pretrained deep convolutional neural network as the base model and applied transfer learning to SAR image classification tasks. They also employed model compression techniques to reduce the number of model parameters, thereby improving the computational efficiency. Rostami et al. [18] presented a deep transfer learning method for few-shot SAR image classification. They pretrained a

model on the source domain and then applied the pretrained model to the target domain with adaptive fine-tuning to improve the classification performance. They also introduced a meta-learning strategy to further enhance the accuracy in the few-shot scenario. Huang et al. [19] investigated deep transfer learning methods for large-scale high-resolution SAR image classification. They proposed a multi-task learning framework to transfer knowledge from the source domain to the target domain. They also strengthened the model's detection capability for small targets by introducing a specific loss function. Huang et al. [20] proposed an SAR target classification method based on deep convolutional neural networks and transfer learning. They employed a pretrained convolutional neural network on a large-scale image dataset as the base model and fine-tuned the network parameters for SAR image classification tasks. Due to the limited annotated samples in SAR image datasets, they also utilized a semi-supervised learning approach to leverage unlabeled samples in the training process. Lu and Li [21] proposed a transfer learning method for ship classification in high-resolution SAR images with a small training dataset. They used a pretrained convolutional neural network on a large-scale image dataset as the base model and fine-tuned the network parameters for ship classification in SAR images. They also augmented the training dataset using data augmentation techniques to enhance the model robustness and generalization. The methods mentioned utilize the concept of transfer learning to adapt pretrained deep convolutional neural networks to specific tasks in remote sensing. The commonality among these methods is that they employ a pretrained model on large-scale image datasets as the base model, and fine-tune or train the model for specific remote sensing tasks. The benefit of this approach is the utilization of learned generic features from large-scale datasets to extract relevant features from remote sensing images, thereby accelerating model training and improving classification performance. These methods also incorporate different strategies to further enhance model performance. For instance, some methods reduce the number of parameters and improve computational efficiency through model compression. Some methods employ meta-learning strategies to adapt to few-shot scenarios in classification tasks. Others enhance model robustness and generalization by introducing specific loss functions or data augmentation techniques [22].

The advancement of deep-learning-based change detection techniques has led to significant research progress [23]. However, existing methods still have some limitations, such as the insufficient fusion of local and global features when processing large-scale data [24], the insufficient accuracy of building change detection in complex environments, and an inability to simultaneously achieve high levels of robustness and generalization [25]. In this paper, we propose a new method for detecting building changes using focal self-attention for local–global interactions in vision transformers, which addresses the issue of overly detailed local and global semantic relationships in self-attention. This method employs the focal self-attention mechanism in the transformer model, which considers local and global feature information. According to the experimental results, the proposed method shows superior performance in building change detection tasks and has practical application value.

2. Materials and Methods

2.1. Dataset

The LEVIR-CD [26] dataset contains 637 pairs of dual-temporal remote sensing images obtained from Google Earth. The images in the dataset have a resolution of 0.5 m and a size of 1024 $\times$ 1024 pixels. Native dual-time remote sensing images were captured between 2002 and 2018 in various areas of Texas, United States. This dataset was divided into smaller patches of 256 $\times$ 256 pixels in order to accommodate GPU memory constraints. For experimental purposes, the dataset is divided into three sets: the training set, the validation set, and the test set. There are 3096 images in the training set, 432 images in the validation set, and 921 images in the test set.

The WHU-CD [27] building dataset covers the Christchurch area of New Zealand, with an area of approximately 450 square kilometers. The dataset consists of three parts:

5201 training images, 744 validation images, and 1487 testing images, with image sizes of 256 × 256 pixels. These images represent the building environment of the region and serve as a benchmark for building detection and related tasks. This dataset provides valuable resources for research in the fields of remote sensing and computer vision, particularly for the development and evaluation of algorithms for building detection and related applications. Figures 1 and 2 show the LEVIR-CD dataset and the WHU-CD dataset, respectively. T1 represents the image before the building change, T2 represents the image after the building change, and GT represents the label, which is the true building change area.

Figure 1. LEVIR-CD dataset. (**a**) The first temporal image. (**b**) The second temporal image. (**c**) The building change label.

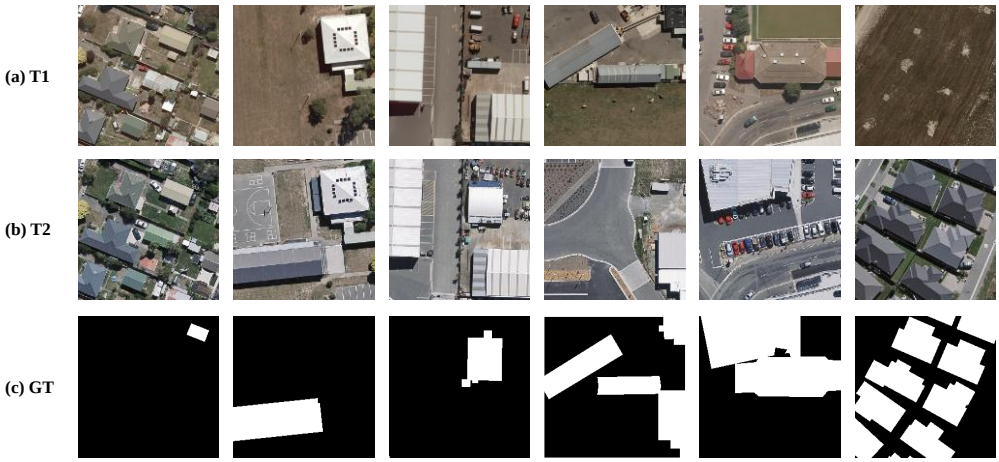

Figure 2. WHU-CD dataset. (**a**)The first temporal image. (**b**) The second temporal image. (**c**) The building change label.

2.2. Network Architecture

We employed the focal self-attention [28] module and a multi-level feature fusion module to identify image change regions. As shown in Figure 3, our network consisted of an encoder and a decoder. The focal self-attention module acts as the encoder, extracting features from the images, while the multi-level feature fusion module acts as the decoder, fusing the extracted features. To begin, the focal self-attention module extracted features from the pre-temporal and post-temporal images, sharing weights across the network. It produced two sets of images, each containing four feature maps of different sizes, corresponding to the same size of feature maps. These sets of feature maps were then fed into the multi-level feature fusion module. In this module, the corresponding feature maps were fused and subjected to convolution to reduce their dimensionality to match that of the

pre-fusion image. Finally, the final output was obtained through feature fusion and feature expansion.

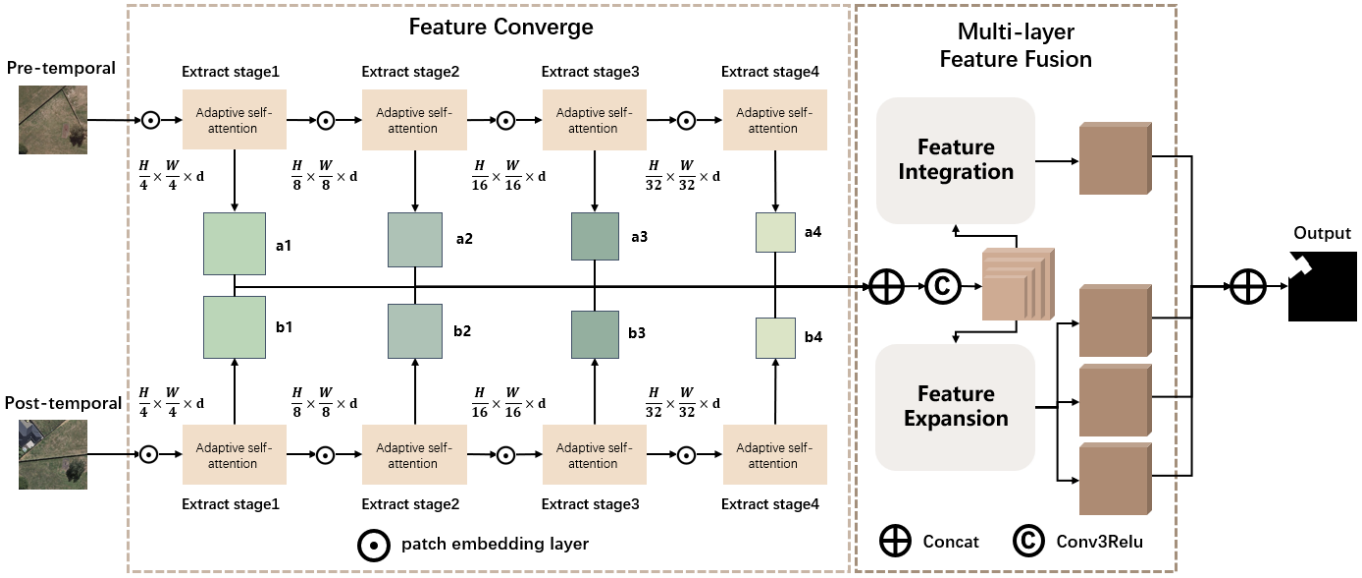

Figure 3. Network architecture diagram.

2.3. Focal Self-Attention Module

In images, the interaction between different regions plays a crucial role, and traditional transformer models capture the interdependencies among these features using self-attention mechanisms. However, they often struggle to effectively capture the interactions between different regions of the image. Therefore, we employed a method called focal self-attention to enhance the interaction between local and global features in the visual transformer model.

In our focal self-attention module, the input image was first processed by a convolutional embedding layer to extract the image's features. The flattened features were then transmitted into the transformer encoder. The encoder contained multiple transformer modules, each consisting of a self-attention layer and a fully connected layer. The feature map was divided into multiple blocks in each self-attention layer. The focal self-attention mechanism calculated the similarity between each block and all other blocks, enabling the selection of specific local regions that shared similarity, thereby focusing on relevant information at those locations. This process can be seen as an enhancement of local–global interaction achieved through local attention mechanisms. Then, a convolutional embedding layer was applied after each stage to decrease the spatial size of the feature map by half, but double the feature dimensionality. This layer was fed into the next transformer module. After multiple layers, the attention-weighted features were fed into multi-head self-attention by the fully connected layer to output the model's prediction results. Additionally, there were two parts to the focal self-attention mechanism. Firstly, it used the self-attention mechanism in the transformer to calculate the interaction between the feature representations of each position and all other positions. Secondly, it introduced an innovative focal mechanism that assigned attention weights to local regions to better capture local–global interaction. The formula is as follows:

$$Att(Q, K, V) = \left(-\frac{QK^T}{\sqrt{d_k}} + M\right)V \quad (1)$$

where Q, K, and V are feature representations of Query, Key, and Value, respectively. d_k represents the feature dimension, and M is a matrix of the same size as Q and K. After

obtaining all the pooled feature maps $\{x_i\}_1^L$, we calculated the queries (Q), keys (K), and values (V) using three linear projection layers: f_q, f_k, and f_v:

$$Q = f_q(x_1), \quad K = \{K_i\}_1^L = f_k(\{x_1, \ldots, x_L\}), \quad V = \{V_i\}_1^L = f_v(\{x_1, \ldots, x_L\}) \tag{2}$$

In the focal self-attention model, we started by extracting the contextual information surrounding each token in the feature map. Tokens within a window of size $s_p \times s_p$ shared the same set of tokens. For the query Q, we extracted correspondingly sized keys and values from the key set K_i and value set V_i, respectively. These keys and values correspond to the tokens surrounding the query within the window. Next, we aggregated the keys and values from all levels to obtain the final key set and value set.

Specifically, for each position, the focal self-attention mechanism calculates its similarity with all other positions and selects a specific local region based on the magnitude of the similarity. This enables the model to focus on position-specific information. We evaluated this method on the LEVIR-CD and WHU-CD datasets, and the results based on various metrics demonstrated significant improvements.

The innovation of the proposed method lies primarily in the use of an Adaptive Focus algorithm to achieve image focusing. Traditional focusing algorithms typically rely on fixed focal lengths or determine the focus position based on image sharpness. In contrast, the Adaptive Focus algorithm dynamically adjusts the focus position by analyzing and evaluating local regions of the image to improve overall image sharpness.

The setting of parameters is also a crucial part. Among them, the focal level is used to specify the level adopted by the focusing algorithm. A lower level means the algorithm focuses more on the overall image sharpness, while a higher level emphasizes the sharpness of local regions. The choice of focusing level depends on specific application scenarios and requirements. Lower levels are suitable for capturing scenes that require overall sharpness, while higher levels are appropriate for emphasizing local details. In the experiment, we set this parameter to 2.

The focal window defines the size of the local window used to evaluate image sharpness. A larger window leads the focusing algorithm to consider more local detail information. However, excessively large windows can increase computational complexity. Therefore, an appropriate window size needs to be balanced based on specific application scenarios and available computational resources. In the experiment, we set this parameter to 7.

The focal factor parameter adjusts the focusing strength in the focusing algorithm. A smaller factor weakens the adjustment effect, while a larger factor enhances it. By properly setting the focal factor, the relationship between global and local sharpness can be balanced according to the image characteristics and requirements. In the experiment, we set this parameter to 2.

The selection and adjustment of these parameters aim to balance the relationship between global and local sharpness and optimize image quality based on specific needs. Through the Adaptive Focus algorithm and proper parameter settings, image sharpness and detail representation can be improved, thereby enhancing image quality. In the experiment, we set the focal level parameter to 2, the focal window to 7, and the focal factor to 2 to simultaneously focus on the overall image sharpness and the sharpness of local regions.

2.4. Multi-Level Feature Fusion Module

To address the issue of multiscale feature fusion in image fusion tasks, this paper proposes a multi-level feature fusion module. As shown in Figure 4, this module first groups the features and then constructs a hierarchical residual structure to fuse the grouped features. This hierarchical residual structure reflects the multiscale nature of the fusion process. The proposed method can be described as follows:

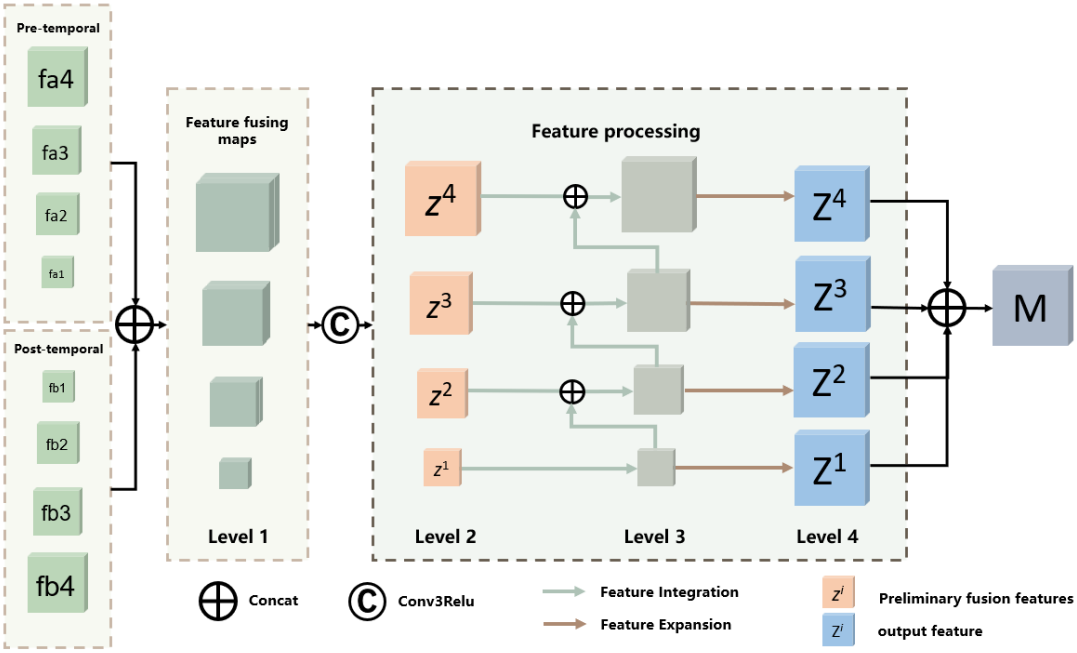

Figure 4. Multi-level feature fusion module.

First, the feature maps of the pre-change image Y_i and the post-change image Y_V are input into a 1×1 convolutional neural network to extract their deep feature maps Y_I^i and Y_V^i. To achieve high-quality fusion of these two feature maps, we utilized a multi-level feature fusion module (MLFFM). This module can fuse the same dimensional images from eight feature maps of four different dimensions. The dimensions of the fused image will be twice those of the original. Then, the concatenated image is convolved to restore the feature dimension to its original size, resulting in four stages. The smallest stage is convolved with Conv3Relu to obtain change1, and so on to obtain four transformed feature maps, namely change1, change2, change3, and change4. Next, the feature fusion method is used to scale the transformed smallest scale feature map (change1) to the same size as change4 using bilinear interpolation. Then, the feature maps are concatenated along the channel dimension and output through Conv3Relu. The entire network incorporates DropBlock regularization, a technique that randomly drops blocks of a certain size at specific locations with a certain probability to mitigate overfitting.

Specifically, for the i-th feature (i ranging from 0 to s, s represents the total size of the focal regions across all levels), we first concatenated the features Y_I^i and Y_V^i, corresponding to the pre- and post-change images according to the channel dimension. Then, we used a 3×3 convolutional layer to alter the dimension and generate the preliminary fused feature z^i. It can be calculated using the following formula:

$$z^i = Conv_i(cat(Y_I^i, Y_V^i)) \tag{3}$$

Then, a progressive strategy was used to gradually fuse different features, and another 3×3 convolution was used to obtain the output feature Z^i. Here is the formula:

$$Z^i = \begin{cases} z^i, i = 1 \\ Conv_F^i(cat(z^i, z^{i-1})), i = 2 \\ Conv_F^i(cat(z^i, Z^{i-1})), 2 < i \leq s \end{cases} \tag{4}$$

where $Conv_i(\cdot)$ and $Conv_F^i(\cdot)$ represent convolutional blocks, including a 3×3 convolutional layer, BN layer, and LeakyRelu activation function. $cat(\cdot)$ represents the operation along the channel dimension. In this paper, we preset s to 3. From the formula, it can be seen that each output Z^i obtained through $Conv_F^i(\cdot)$ potentially contains the preliminary

fusion feature z^i. This allows more detailed information and global information to fully interact and fuse.

Finally, the intermediate fusion results were concatenated along the channel dimension to obtain the fusion feature map M. Here is the formula:

$$M = cat(Z^1, Z^2, \ldots Z^i) \tag{5}$$

In this way, the multi-level feature fusion module design was completed, and the entire module could fully utilize feature information of different scales, allowing for the high-quality fusion of the extracted feature maps.

2.5. Loss Function

In most deep learning tasks, the ratio of changed areas is typically smaller than that of unchanged areas, leading to a class imbalance in the image. In this case, a loss function can be used to improve feature extraction accuracy. The *BCE* (Binary Cross-Entropy) loss function [29] is a commonly used loss function for binary classification, and the formula is as follows:

$$LOSS_{BCE} = -\{yIog[p(x) + (1-y)\log(1-p(x))]\} \tag{6}$$

Here, *p(x)* is the model output and y is the ground truth label.

The Binary Cross-Entropy (*BCE*) loss function calculates the prediction error of the model in binary classification problems, aiding the parameter adjustment to improve accuracy. However, the *BCE* loss function may not be as effective in multi-class tasks. The *DICE* (Sørensen Dice) loss function is a commonly used method for calculating the loss in image segmentation tasks. The formula for this loss function is as follows:

$$LOSS_{DICE} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \tag{7}$$

Here, the label value and predicted value of pixel i are, respectively, represented by y_i and $\hat{y}_i$, while N represents the pixels' total number.

To enhance the model's accuracy, the *DICE* loss function facilitates the learning of precise boundary information. However, it presents challenges in detecting small targets, as pixel prediction errors for such targets can lead to substantial variations in the *DICE* loss function, resulting in drastic changes in gradients. Therefore, we employed a combined loss function of *BCE* and *DICE* to improve the overall performance of the model. In comparison to the individual *BCE* and *DICE* loss functions, the combined loss function comprehensively incorporated pixel-level predictions and image-level predictions. Regarding the model's generalization capability, it mitigated the issue of model overfitting. Regarding image fusion, it additionally extracted edge features and internal information from the image, thereby enhancing the model's accuracy. The following is the formula:

$$L(C, G) = L_{BCE}(C, G) + L_{DICE}(C, G) \tag{8}$$

3. Experiments and Results

3.1. Experimental Settings

We implemented our proposed method based on deep learning and the Pytorch framework. The Adam gradient optimization algorithm was used with a learning rate of 0.0035, and the iteration was set to 500. The GPU used was NVIDIA Tesla A100 with a batch size of 16.

3.2. Evaluation Metrics

This article employed Precision, Recall, F1-score, and IOU as evaluation metrics. Precision represents the ratio of correctly identified change regions to the total detected change regions. Recall reflects the proportion of true positive examples correctly identified. F1-score is a metric that combines Precision and Recall using their harmonic mean, providing a comprehensive assessment of the classifier's performance. IOU is employed to measure the accuracy in detecting the target location. These metrics provide a comprehensive evaluation of the algorithm's performance. The formulas for calculating these metrics are presented below.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

$$\text{IOU} = \frac{TP}{TP + FP + FN} \tag{12}$$

3.3. Comparison Methods

The following sections provide explanations of seven other representative methods.

The FC-EF [30] is a U-Net-based method that takes dual-temporal image connections as input and uses skip connections to achieve feature mapping.

The FC-Siam-conc [30] employs a Siamese network with shared weights to extract multiscale features and combines cascading operations to merge these features through skip connections.

The FC-Siam-diff [30] is a change detection approach similar to FC-Siam-conc, but it utilizes absolute difference instead of concatenation.

The CDNet [31] consists of four blocks of compression and four blocks of expansion. The compression blocks obtain image features, the expansion blocks optimize detection results, and the softmax block classifies each pixel.

The LUNet [32] integrates fully convolutional LSTM modules on top of each encoding layer of a U-Net-like deep architecture, computing the temporal relationship of feature vectors of different resolutions without downsampling or flattening.

The IFNet [33] utilizes a deep supervised difference discrimination network for detecting changes, using a spatio-temporal attention strategy to merge information from different scales.

The BITNet [34] is a transformer-based network that adds a Siamese-style marker to obtain semantic markers of dual-temporal images, enhancing features' global contextual effectiveness.

3.4. Visual Comparison Results

We conducted experiments using a set of eight different algorithms on two datasets. Figures 5 and 6, respectively, present the three sets of experimental results obtained for each dataset. Each set of experimental results comprises two pairs of dual-temporal remote sensing images, one labeled image, seven comparative experimental results, and the results obtained from our proposed method.

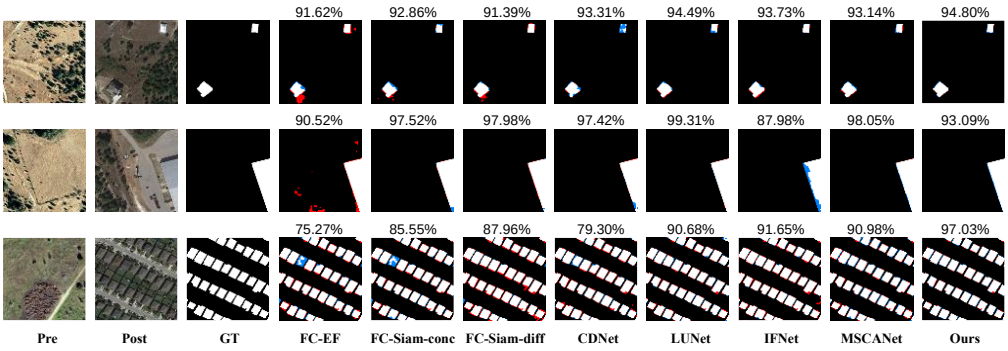

Figure 5. Image results of various models in the LEVIR-CD dataset. The numbers above the images represent F1-score scores.

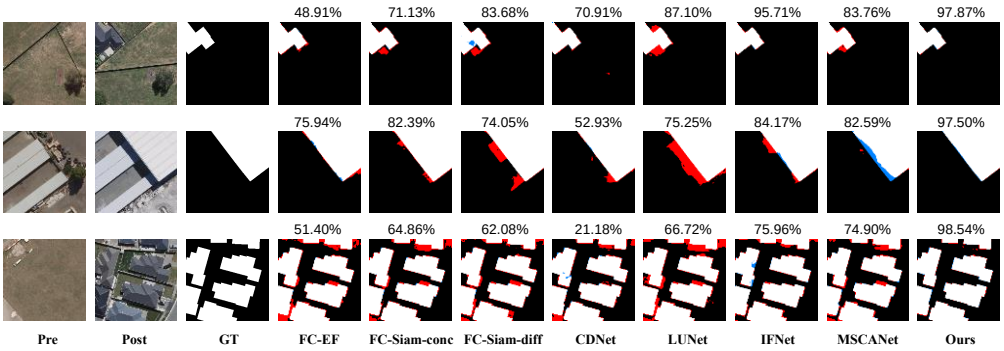

Figure 6. Image results of various models in the WHU-CD dataset. The numbers above the images represent F1-score scores.

Moreover, the article visually demonstrates the superior performance of our proposed method compared to other existing methods on the LEVIR-CD dataset, as depicted in Figure 5. In particular, owing to the utilization of the focal self-attention module and multi-level feature fusion module, our method achieves a more accurate detection of building changes, as observed in the first row of Figure 5, which represents areas with small building changes. In the second row of Figure 5, corresponding to areas with medium building changes, FC-EF exhibits sensitivity to changes in regions with high spectral brightness coefficients, leading to notable detection errors, whereas IFNet struggles to accurately detect the change area. Conversely, our proposed method demonstrates the accurate detection of building changes. In the third row of Figure 5, corresponding to areas with dense building changes, our proposed method achieves a lower missed detection rate compared to FC-Siam-conc and FC-EF, as well as a lower false detection rate compared to FC-Siam-diff.

To quantitatively measure the visual results, we evaluated the performance of different methods on the LEVIR-CD dataset using the F1-Score, as shown in Figure 5. It can be observed that the comparative methods achieved higher F1-Score scores in the first two instances, but experienced a significant decline in the final F1-Score. In contrast, our proposed method maintained the high F1-Score scores from the previous instances and demonstrated a substantial improvement in the final performance, thus validating the effectiveness of our model. Various methods' results from the WHU-CD dataset are shown in Figure 6. In the first two experiments of this dataset, FC-EF, IFNet, and our method can identify the real change contours of buildings. However, all methods show false detections when dealing with dense building changes. This is mainly because they are not sensitive to the shadows cast by buildings. They may identify the shadows as buildings. In the third row, FC-Siam-diff, FC-Siam-conc, FC-EF, LUNet, and BITNet are not sensitive to the color distinction between buildings and the ground, and may misidentify the ground as buildings. The proposed method incorporates a focal self-attention module and a multi-level feature fusion module to enhance the extraction of effective features in images

and strengthen the differential features between buildings and non-buildings, thereby reducing the misclassification of buildings as non-buildings or non-buildings as buildings. Similarly, we conducted a quantitative evaluation on the WHU-CD dataset, as shown in Figure 6. It can be observed that the WHU-CD dataset is more challenging, as the majority of methods achieved low F1-score scores across all three instances. Only the IFNet method surpassed a 90% F1-score in the first instance. In contrast, our proposed method exhibited excellent performance, with F1-score scores exceeding 95% in all three instances. This performance significantly outperformed the comparative methods, further validating the reliability of the visual analysis results discussed earlier. These findings demonstrate the strong performance of our model in handling complex environments. In summary, the proposed method has a comprehensive extraction of effective features and is relatively efficient in performance. It can accurately identify buildings and non-buildings. It also has an excellent detection ability for some high discrimination changes, and the detection results are relatively complete.

3.5. Quantitative Analysis

To validate the method's stability and effectiveness, we conducted a quantitative analysis on the LEVIR-CD dataset. Results are presented in Table 1. The proposed method achieved a precision of 4.98% higher than the current optimal BITNet method, while maintaining a comparable recall rate. Moreover, the proposed method also outperformed existing methods on F1-score, which comprehensively evaluates the performance by considering both precision and recall. Additionally, compared with other methods, the proposed method produced the best results for the IOU performance metric. The data presented in Table 1 indicate the superiority of our proposed method over other existing approaches in terms of accuracy and completeness in detecting building change and delineating model boundaries. Specifically, thanks to the focal self-attention module and multi-level feature fusion module, the proposed method can detect the range of building changes more accurately in datasets with small building change areas. In datasets with moderate building change areas, FC-EF was sensitive to changes in areas with large spectral brightness coefficients in the image and had significant detection errors, while IFNet failed to detect the change area well. In contrast, our proposed method demonstrates the accurate detection of building changes. In datasets with dense building changes, compared to FC-EF and FC-Siam-conc, this method has a lower miss detection rate, and compared to FC-Siam-diff it has a reduced rate of false positives. Considering all these factors, our method achieved the highest precision, F1-score, and IOU.

Table 1. Evaluation metrics for the LEVIR-CD dataset. The best results are highlighted in red.

Method	Precision (%)	Recall (%)	F1-Score (%)	IOU (%)
FC-EF	79.91	82.84	81.35	68.56
FC-Siam-conc	81.84	83.55	82.68	70.48
FC-Siam-diff	78.60	89.30	83.61	71.84
CDNet	84.21	87.10	85.63	74.87
LUNet	85.69	90.99	88.73	79.75
IFNet	85.37	90.24	87.74	78.16
BITNet	87.32	91.41	89.32	80.70
Ours	92.30	90.96	91.62	84.54

For the WHU-CD dataset, we also performed a quantitative analysis. Table 2 shows the result. The proposed method exhibited an excellent detection performance, with three metrics—precision, F1-score, and IOU—outperforming the current optimal methods. In addition, the proposed method also achieved a high F1-score, while maintaining high precision. The experimental results indicate that in datasets with small building change areas, the proposed method exhibited a lower false detection rate than FC-Siam-conc and LUNet. In datasets with moderate building change areas, LUNet had significant

detection errors and BITNet failed to detect the change area well. By contrast, the proposed method showed advantages in detection accuracy. In datasets with dense building changes, current existing methods were not sensitive to the color distinction between buildings and the ground, resulting in serious false detection problems, while the proposed method, combined with the multi-level feature fusion module and focal self-attention module, better utilized feature information at different levels and focused more on important features of the target object, thereby improving the accuracy of object detection and classification. In terms of the accuracy of detecting changes in dense buildings, our method achieved significant advancements.

Table 2. Evaluation metrics for the WHU-CD dataset. The best result is highlighted in red.

Method	Precision (%)	Recall (%)	F1-Score (%)	IOU (%)
FC-EF	70.43	92.31	79.90	66.53
FC-Siam-conc	63.80	91.81	75.28	60.36
FC-Siam-diff	65.98	94.30	77.63	63.44
CDNet	81.75	88.69	85.08	74.03
LUNet	66.32	93.06	77.45	63.19
IFNet	86.51	87.69	87.09	77.14
BITNet	82.35	92.59	87.17	77.26
Ours	90.86	88.08	89.45	80.91

We also visually demonstrate the advantages of this method in detecting building changes more accurately and with more complete model boundaries compared to other existing methods, as shown in the visualization result curve in Figure 7. The curve shows that this method has some advantages in the IOU metric and has higher accuracy in detecting the location of building changes compared to current existing methods. Although the proposed method has some shortcomings in recall compared to current methods, it achieved a high F1-score without sacrificing precision, ensuring the model's comprehensive performance.

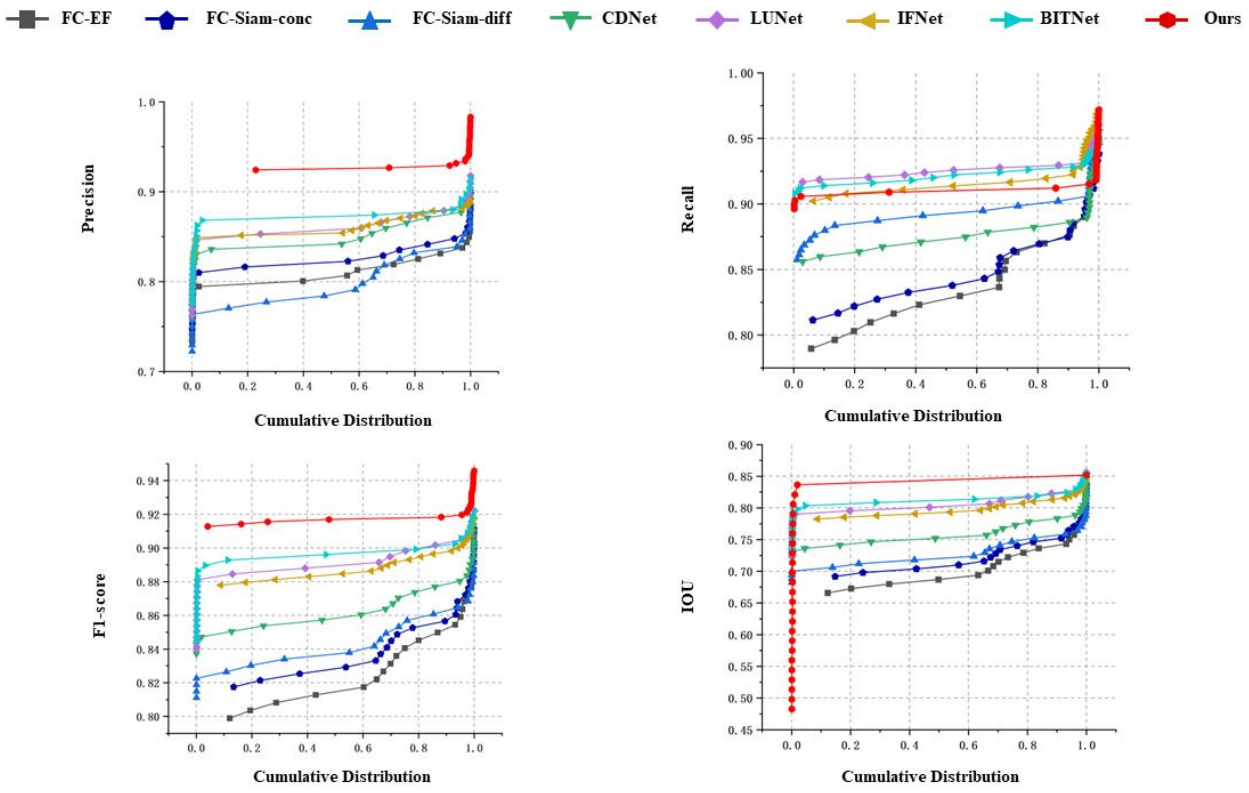

Figure 7. Curve graph of evaluation metrics.

In the binary building change detection task, there was a significant imbalance between positive and negative samples (the number of negative samples was much larger than the number of positive samples). Inspired by [2,23], we added an analysis of the precision–recall curve to better represent the performance of different algorithms in the context of imbalanced data.

As shown in Figure 8, we randomly selected five sets of images from the LEVIR-CD dataset and calculated the precision and recall values for each method. We then plotted the precision–recall curve based on these values. We observed that recall and precision showed a positive correlation in general. The recall values for FC-EF and FC-Siam-conc were concentrated below 0.94 because their early fusion strategy led to the loss of some feature information. Thanks to the effectiveness of the multi-scale fusion strategy, IFNet and our method performed relatively well. In particular, our multi-level feature fusion module could fuse features from different scales, resulting in promising quantitative performance metrics.

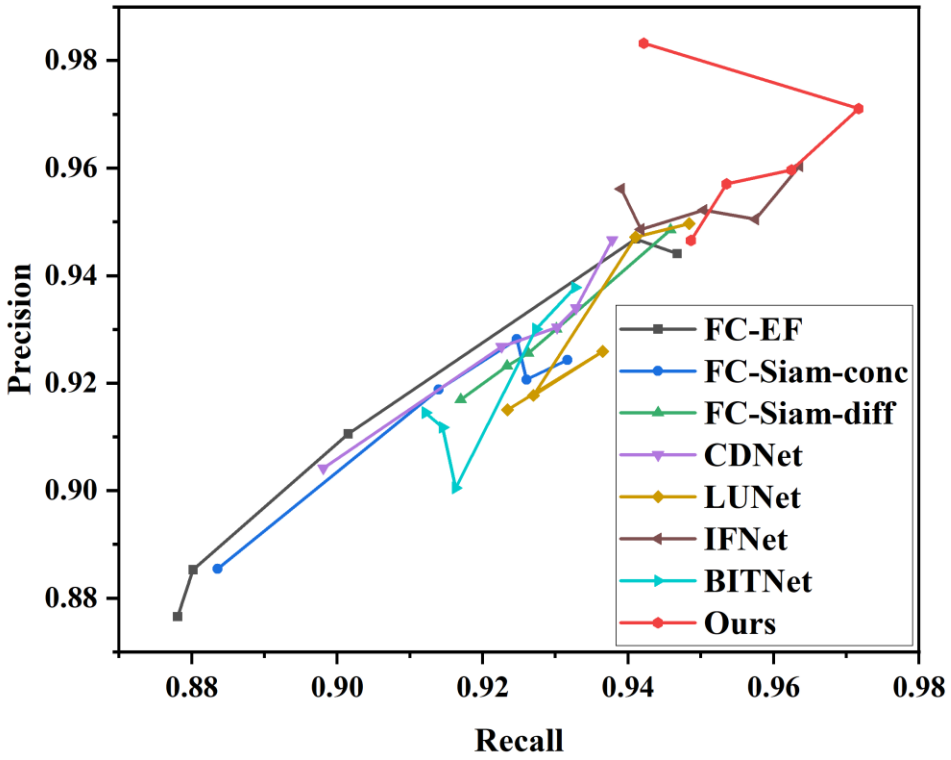

Figure 8. Precision–recall curve from the LEVIR-CD dataset. Note that we randomly selected five points for demonstration.

3.6. Ablative Analysis

In addition, we conducted ablative experiments to further investigate the effectiveness of our proposed multi-level feature fusion module, as shown in Table 3. We removed the multi-level feature fusion module and observed the corresponding changes in performance metrics. For the LEVIR-CD dataset, it can be seen that with the removal of the multi-level feature fusion module, almost all the metrics scores decreased. The precision metric decreased by 1.57%, recall decreased by 2.07%, and the F1-score decreased by 1.82%. For the WHU-CD dataset, the precision metric showed a slight improvement, but the recall metric decreased by 2.72% and the F1-score decreased by 1.32%. This is because without the multi-level feature fusion module, the model is unable to effectively integrate global feature information, resulting in a decrease in recognition performance. By comparing the results of these ablative experiments with the performance of the complete model, we quantified the contribution of the multi-level feature fusion module, which helped

evaluate its impact on overall performance. This validation confirmed the effectiveness of our proposed method and its role in improving detection accuracy.

Table 3. The ablative experimental results on the LEVIR-CD dataset and WHU-CD dataset. Note that "MLFF" represents the multi-level feature fusion module.

Dataset	MLFF	Precision (%)	Recall (%)	F1-Score (%)	IOU (%)
LEVIR-CD	×	90.73	88.89	89.80	81.49
	√	92.30	90.96	91.62	84.54
WHU-CD	×	91.08	85.36	88.13	78.78
	√	90.86	88.08	89.45	80.91

4. Discussion

This paper proposes a focal self-attention module and a multi-level feature fusion method for detecting building changes in remote sensing images. This method comprehensively incorporates both local and global feature information and emphasizes key areas during image feature processing, leading to an improved detection accuracy. In comparative experiments with seven other methods, our method showed good results. It achieved an accuracy of 92.3% on the LEVIR-CD dataset, which is 4.98% higher than the current best method, BITNet, and basically did not sacrifice recall rate. Moreover, our method achieved the highest scores in terms of F1-score and IOU. Likewise, on the WHU-CD dataset, our method outperformed the compared methods in the precision, F1-score, and IOU metrics. In comparison with other existing methods, this method provided better building change detection accuracy and more complete model boundaries, proving its more comprehensive performance.

Among the comparative methods, FC-EF was sensitive to changes in regions with high spectral brightness coefficients, leading to significant detection errors. IFNet could not accurately detect change areas in certain cases. FC-EF, FC-Siam-diff, FC-Siam-conc, LUNet, and BITNet were not sensitive in distinguishing building and ground colors, which could result in the ground being misidentified as a building. In dense building areas, almost every method produces false detections because they were not sensitive to the shadows produced by buildings, which were also identified as buildings. The proposed method incorporated a focal self-attention module and a multi-level feature fusion module, enhancing the extraction of effective features in images and strengthening the differential features between buildings and non-buildings, thus reducing false detections. The proposed focal self-attention module calculated the similarity between each position's features and those of all other positions and assigned attention weights to local regions, selecting local regions based on their similarity to better facilitate local–global interactions and extract high-level semantic information. The proposed multi-level fusion module used deep convolution and self-attention to comprehensively extract change information from shallow to deep levels to locate the actual change position, thereby extracting effective features more comprehensively and improving change detection accuracy.

To evaluate the generalization ability of our proposed methodology, we conducted cross-testing on two datasets, as shown in Table 4. We tested the weights trained on the WHU-CD dataset on the LEVIR-CD dataset, and vice versa. This setup allowed us to assess the performance of the model on different datasets and infer its generalization ability. However, we observed performance discrepancies when conducting cross-testing. Specifically, when testing the weights trained on the WHU-CD dataset on the LEVIR-CD dataset, we observed a precision of 45.70%, a recall of 4.78%, an F1-score of 8.60%, and an IOU of 4.53%. Similarly, when testing the weights trained on the LEVIR-CD dataset on the WHU-CD dataset, the precision, recall, F1-score, and IOU scores were 58.98%, 53.91%, 56.33%, and 39.21%, respectively.

Table 4. The cross-validation experimental results on the LEVIR-CD dataset and WHU-CD dataset.

Dataset	Precision (%)	Recall (%)	F1-Score (%)	IOU (%)
LEVIR-CD	45.70	4.78	8.60	4.53
WHU-CD	58.98	53.91	56.33	39.21

These results indicate that there is room for improvement in the generalization ability of our proposed methodology across different datasets. The main reason for this is the distinct characteristics and data distributions between the WHU-CD and LEVIR-CD datasets, making it challenging for the model to adapt to the varying patterns and feature representations of different datasets. Hence, we acknowledge the limitations of our model in handling different datasets. While our model achieved a satisfactory performance on specific datasets, its limited generalization ability was a significant drawback. We need to further enhance the design and training strategies of the model to improve its adaptability and generalization across different datasets. Possible improvements include incorporating data augmentation techniques, utilizing diverse training datasets, and adjusting the model's complexity to cater to a broader range of building change detection scenarios.

5. Conclusions

In this paper, we adopted a focal self-attention module and a multi-level feature fusion method to process building change detection and validated our approach on two datasets. The results show that our method surpassed the SOTA system and achieved an excellent performance. The proposed method for detecting building changes in remote sensing images also exhibited a higher robustness than existing methods. However, our method still has room for improvement compared to other methods. The model needs further optimization to handle more complex scenarios and harsher environments. Overall, this method still has significant potential and can provide valuable references for research and applications in remote sensing for detecting building changes.

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

Funding: This research was funded by the National Natural Science Foundation of China (Nos. 41601443); the Scientific Research Foundation for Doctoral Programs of Hubei University of Technology (BSQD2020056); the Science and Technology Research Project of the Education Department of Hubei Province (B2021351); the Natural Science Foundation of Hubei Province (2022CFB501); and the University Student Innovation and Entrepreneurship Training Program Project (202210500028).

Data Availability Statement: The LEVIR-CD dataset that supported the findings of this study is available in Google Earth at https://justchenhao.github.io/LEVIR/. The WHU-CD dataset that supported the findings of this study is available at http://gpcv.whu.edu.cn/data/building_dataset.html.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, C.; Ye, Z.; Mei, L.; Yang, W.; Hou, Y.; Shen, S.; Ouyang, W.; Ye, Z. Progressive Context-Aware Aggregation Network Combining Multi-Scale and Multi-Level Dense Reconstruction for Building Change Detection. *Remote Sens.* **2023**, *15*, 1958. [CrossRef]
2. Islam, K.A.; Uddin, M.S.; Kwan, C.; Li, J. Flood detection using multi-modal and multi-temporal images: A comparative study. *Remote Sens.* **2020**, *12*, 2455. [CrossRef]
3. Wang, D.C.; Zhao, F.; Wang, C.; Wang, H.Y.; Zheng, F.J.; Chen, X.N. Y-Net: A Multiclass Change Detection Network for Bi-temporal Remote Sensing Images. *Int. J. Remote Sens.* **2022**, *43*, 565–592. [CrossRef]
4. Kwan, C. Methods and challenges using multispectral and hyperspectral images for practical change detection applications. *Information* **2019**, *10*, 353. [CrossRef]
5. Yang, W.; Xu, C.; Mei, L.Y.; Yao, Y.X.; Liu, C. LPSO: Multi-Source Image Matching Considering the Description of Local Phase Sharpness Orientation. *IEEE Photonics J.* **2022**, *14*, 7811109. [CrossRef]

6. Javed, A.; Jung, S.; Lee, W.H.; Han, Y. Object-Based Building Change Detection by Fusing Pixel-Level Change Detection Results Generated from Morphological Building Index. *Remote Sens.* **2020**, *12*, 2952. [CrossRef]
7. Guo, X.P.; Meng, L.Y.; Mei, L.Y.; Weng, Y.Y.; Tong, H.Q. Multi-focus image fusion with Siamese self-attention network. *IET Image Process.* **2020**, *14*, 1339–1346. [CrossRef]
8. Gong, M.G.; Zhan, T.; Zhang, P.Z.; Miao, Q.G. Superpixel-Based Difference Representation Learning for Change Detection in Multispectral Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2658–2673. [CrossRef]
9. Seo, D.K.; Kim, Y.H.; Eo, Y.D.; Park, W.Y.; Park, H.C. Generation of Radiometric, Phenological Normalized Image Based on Random Forest Regression for Change Detection. *Remote Sens.* **2017**, *9*, 1163. [CrossRef]
10. Canty, M.J.; Nielsen, A.A. Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation. *Remote Sens. Environ.* **2008**, *112*, 1025–1036. [CrossRef]
11. Gao, W.; Sun, Y.; Han, X.; Zhang, Y.; Zhang, L.; Hu, Y. AMIO-Net: An Attention-Based Multiscale Input–Output Network for Building Change Detection in High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2079–2093. [CrossRef]
12. Peng, D.F.; Bruzzone, L.; Zhang, Y.J.; Guan, H.Y.; Ding, H.Y.; 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]
13. Xiao, P.F.; Yuan, M.; Zhang, X.L.; Feng, X.Z.; Guo, Y.W. Cosegmentation for Object-Based Building Change Detection from High-Resolution Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 1587–1603. [CrossRef]
14. Zhang, Y.; Deng, M.; He, F.; Guo, Y.; Sun, G.; Chen, J. FODA: Building Change Detection in High-Resolution Remote Sensing Images Based on Feature–Output Space Dual-Alignment. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8125–8134. [CrossRef]
15. Zhang, J.; Pan, B.; Zhang, Y.; Liu, Z.; Zheng, X. Building Change Detection in Remote Sensing Images Based on Dual Multi-Scale Attention. *Remote Sens.* **2022**, *14*, 5405. [CrossRef]
16. Zhou, J.; Kwan, C.; Ayhan, B.; Eismann, M.T. A novel cluster kernel RX algorithm for anomaly and change detection using hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6497–6504. [CrossRef]
17. Zhong, C.; Mu, X.; He, X.; Wang, J.; Zhu, M. SAR target image classification based on transfer learning and model compression. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 412–416. [CrossRef]
18. Rostami, M.; Kolouri, S.; Eaton, E.; Kim, K. Deep transfer learning for few-shot SAR image classification. *Remote Sens.* **2019**, *11*, 1374. [CrossRef]
19. Huang, Z.; Dumitru, C.O.; Pan, Z.; Lei, B.; Datcu, M. Classification of large-scale high-resolution SAR images with deep transfer learning. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 107–111. [CrossRef]
20. Huang, Z.; Pan, Z.; Lei, B. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. *Remote Sens.* **2017**, *9*, 907. [CrossRef]
21. Lu, C.; Li, W. Ship classification in high-resolution SAR images via transfer learning with small training dataset. *Sensors* **2018**, *19*, 63. [CrossRef] [PubMed]
22. Kwan, C.; Chou, B.; Hagen, L.; Perez, D.; Shen, Y.; Li, J.; Koperski, K. Change detection using Landsat and Worldview images. In *Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imagery XXV, 2019*; SPIE: Bellingham, WA, USA, 2019; pp. 328–342.
23. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
24. Mei, L.Y.; Yu, Y.L.; Shen, H.; Weng, Y.Y.; Liu, Y.; Wang, D.; Liu, S.; Zhou, F.L.; Lei, C. Adversarial Multiscale Feature Learning Framework for Overlapping Chromosome Segmentation. *Entropy* **2022**, *24*, 522. [CrossRef] [PubMed]
25. Xiao, J.; Guo, H.; Zhou, J.; Zhao, T.; Yu, Q.; Chen, Y.; Wang, Z. Tiny object detection with context enhancement and feature purification. *Expert Syst. Appl.* **2023**, *211*, 118665. [CrossRef]
26. Chen, H.; Shi, Z.W. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
27. 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.* **2019**, *57*, 574–586. [CrossRef]
28. Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; Gao, J. Focal self-attention for local-global interactions in vision transformers. *arXiv* **2021**, arXiv:2107.00641.
29. Li, Q.Y.; Zhong, R.F.; Du, X.; Du, Y. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622519. [CrossRef]
30. 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; IEEE: New York, NY, USA, 2018; pp. 4063–4067.
31. Alcantarilla, P.F.; Stent, S.; Ros, G.; Arroyo, R.; Gherardi, R. Street-view change detection with deconvolutional networks. *Auton. Robot.* **2018**, *42*, 1301–1322. [CrossRef]

32. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7651–7668. [CrossRef]
33. 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]
34. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]