remote sensing

MDPI

Article

A CNN- and Transformer-Based Dual-Branch Network for Change Detection with Cross-Layer Feature Fusion and Edge Constraints

Xiaofeng Wang, Zhongyu Guo and Ruyi Feng *

School of Computer, China University of Geosciences, Wuhan 430074, China; wxf@cug.edu.cn (X.W.); guozhongyu@cug.edu.cn (Z.G.)
* Correspondence: fengry@cug.edu.cn

Abstract: Change detection aims to identify the difference between dual-temporal images and has garnered considerable attention over the past decade. Recently, deep learning methods have shown robust feature extraction capabilities and have achieved improved detection results; however, they exhibit limitations in preserving clear boundaries for the identified regions, which is attributed to the inadequate contextual information aggregation capabilities of feature extraction, and fail to adequately constrain the delineation of boundaries. To address this issue, a novel dual-branch feature interaction backbone network integrating the CNN and Transformer architectures to extract pixel-level change information was developed. With our method, contextual feature aggregation can be achieved by using a cross-layer feature fusion module, and a dual-branch upsampling module is employed to incorporate both spatial and channel information, enhancing the precision of the identified change areas. In addition, a boundary constraint is incorporated, leveraging an MLP module to consolidate fragmented edge information, which increases the boundary constraints within the change areas and minimizes boundary blurring effectively. Quantitative and qualitative experiments were conducted on three benchmarks, including LEVIR-CD, WHU Building, and the xBD natural disaster dataset. The comprehensive results show the superiority of the proposed method compared with previous approaches.

Keywords: change detection; Transformer; feature fusion; edge constraints; cross-layer

Citation: Wang, X.; Guo, Z.; Feng, R. A CNN- and Transformer-Based Dual-Branch Network for Change Detection with Cross-Layer Feature Fusion and Edge Constraints. *Remote Sens.* **2024**, *16*, 2573. https://doi.org/10.3390/rs16142573

Academic Editor: Pedro Melo-Pinto

Received: 6 June 2024
Revised: 4 July 2024
Accepted: 11 July 2024
Published: 13 July 2024

1. Introduction

Remote sensing image change detection aims to identify pixel-level changes between dual-temporal images, which is a crucial research focus within the fields of pattern recognition and computer vision [1]. Presently, it has extensive application in diverse domains, including monitoring natural disasters [2], tracking urban expansion [3], analyzing agricultural changes [4], and studying environmental evolution [5].

Before the ascent of deep learning, traditional change detection methods primarily involved comparing pixels. These methods typically required the design of artificial features to depict pixel disparities, thus depending on considerable expertise and experience [6]. Moreover, it is difficult to accurately distinguish change from non-change areas when confronted with occlusions or complex scene changes. With the progress of technology, deep learning has found widespread application in many fields. By constructing multi-layer neuron structures, these technologies learn some abstract features of images, which diminishes reliance on expert knowledge and makes them very suitable for the field of remote sensing imaging [7,8]. While the detection results have been optimized to some extent, the deep-learning-based method struggles to effectively capture features of the changed area, indicating that there remains potential for further improvement in the detection results. The current methods of change detection based on deep learning have become mainstream;

they can be categorized into three forms based on different tasks: pixel-level, object-level, and scene-level detection [9].

Pixel-level change detection uses independent pixels as detection units and extracts change information by analyzing pixel differences with pixel-by-pixel operations, which is commonly employed in the initial stages of change detection. Typical approaches include the differential technique, the ratio method, and other direct pixel comparison methods [10]. But these methods often fail to use image features effectively, which limits accuracy. In response to these limitations, scholars have proposed statistics-based detection methods, such as change vector analysis, principal component analysis, and texture-based analysis, as well as post-classification comparison methods that compare pixels after classification. However, these methods tend to rely on fixed features, so they are susceptible to environmental changes in images, such as lighting variations or shadows, resulting in poor performance in actual scenarios [11]. Subsequently, machine learning methods, such as artificial neural networks, support vector machine, decision tree, and random forest, have gained traction in change detection. Compared with traditional methods, machine learning approaches demonstrate significant improvements in accuracy. After the widespread adoption of deep learning, many researchers have begun treating pixel-level change detection as a semantic segmentation problem, applying models from the segmentation field to change detection tasks [12–15].

Object-level change detection utilizes various feature information from dual-temporal images and segments objects within images [16]. As a key to object-level change detection, object generation necessitates ensuring the consistency of object boundaries at different times. In the early stages, traditional algorithms, such as the Robert operator, the Laplacian operator, or region segmentation algorithms, were commonly employed for image segmentation [17]. However, these methods fell short in obtaining object boundaries. Currently, there are three approaches to object generation: single-temporal segmentation boundary, multi-temporal segmentation, and combined segmentation [18]. A single boundary is applied across all temporal intervals in the first approach, which avoids the need for complex cross-temporal object matching and alignment. Although the single-temporal method entails lower overall computational complexity, the detection results are not sufficiently accurate. In contrast, the multi-temporal segmentation approach yields finer segmentation objects by using boundary superimposition, resulting in greater robustness. The combined segmentation method involves multi-temporal remote sensing image bands, addressing the limitations of single-temporal approaches and enhancing detection accuracy. However, this method also introduces the challenge of computational complexity. In sum, object-level change detection methods focus not only on changes in the pixel value but also on changes in objects, which contain more semantic information.

Scene-level change detection employs multi-temporal remote sensing images as units to assess changes across all pixels at the same time [19]. It integrates both local and global information, effectively reducing the influence of noise. However, scene-level methods focus on the whole scene, which requires substantial computer memory when processing large scenes. In 2019, an enhanced U-Net++ network was introduced by Peng et al. [20]. By combining a fully convolutional neural network and the U-Net structure, the model can not only adapt to images of any size for end-to-end training but also improve detection accuracy. Building upon Peng et al.'s work, Lin proposed a new way to cut remote sensing images into regular image blocks and input them into the network to judge changes, which represents a new direction for scene-level change detection [21]. Subsequently, by optimizing the size of image blocks, Li et al. proposed a model with further enhanced performance and computational efficiency in scene-level change detection [22].

While object-level and scene-level change detection offer higher detection accuracy, they require predefined object or scene definitions, which requires considerable manpower and time. On the other hand, pixel-level change detection relies on a simpler data source and has more flexible application scenarios. Consequently, researchers' attention is currently largely directed towards the simpler pixel-level change detection approach, on which

we also focused in this study. At present, pixel-level change detection models have three stages: feature extraction, feature fusion, and upsampling. In feature extraction and feature fusion, the utilization of multi-scale features from dual-temporal images has been demonstrated to be an effective method for predicting subtle changes and enhancing change detection accuracy [23]. Therefore, scholars have tried to combine a variety of excellent feature extraction modules to extract multi-scale feature information and fuse it to improve accuracy in pixel-level change detection [24–26]. There are two main strategies: One approach involves the use of Transformer instead of a CNN as the backbone network to extract better feature information. Although Transformer-based models have a larger receptive field and can better grasp the change region, the local detail information processing ability for the edge information of the change region is limited, with the predicted change region usually presenting blurred edges. The second approach is based on the use of a U-Net structure to fuse contextual feature information [27–29]. This type of structure can integrate multi-level contextual information; nevertheless, it is hindered by the unsophisticated upsampling method, which restricts the model's learning capabilities. Therefore, at present, these two strategies cannot fully integrate image information or generate sufficiently accurate change maps. This makes these models prone to false and missed detection and also to presenting fuzziness in the edge area of the change map.

To address the issue of inadequate feature representation and extraction in detection models and to mitigate edge blurring to provide distinct predictions of the boundary of change areas, a change detection model incorporating cross-layer feature fusion and edge constraints is proposed. The primary contributions of our study can be outlined as follows:

1. A fusion network based on a CNN and Transformer was designed as a feature extractor. In the feature extraction stage, the CNN structure is used to extract local feature information, and Transformer is used to extract global feature information. Then, the features are fused by using the spatial feature interaction module and feature fusion module to promote the correlation between local and global information, optimize the model objects and missed detection, and help improve accuracy.
2. The addition of a boundary constraint module based on the MLP structure allows segmented edge information to be integrated into the boundary of the constrained change area in the feature map. In order to improve the learning ability of the model, the Bilinear and Pixel Shuffle methods are used to upsample the spatial and channel dimensions, respectively.

The remainder of this paper is organized as follows: Section 2 presents an overview of the change detection literature. Section 3 provides the overall details of the model design. Section 4 introduces the experimental part of our proposed model, and finally, Section 5 summarizes our findings and future work.

2. Related Works

2.1. Classical Change Detection Methods

In the early stages of change detection, various methods were established to generate difference maps by comparing the corresponding pixels in dual-temporal images and then form the final change map by using threshold segmentation or region-growing techniques. However, the quality of these change maps was poor, especially in occluded areas. Thanks to the robust feature representation and nonlinear modeling capabilities of deep learning, change detection results have significantly improved. Deep-learning-based change detection directly employs pixel-level classification maps as results with an end-to-end framework, with convolutional neural networks (CNNs) being the most prevalent approach.

In CNN-based change detection, the relationship between dual-temporal images is established with feature mapping functions, which reduces the influence of noise compared with traditional methods. However, in order to achieve better results, radiation and geometric corrections need to be applied to early methods, which fall short compared with end-to-end frameworks. In 2018, an end-to-end 2D CNN change detection model for

hyperspectral images was developed by Wang et al. [30]; it extracts spectral and spatial feature information from dual-temporal hyperspectral images and fuses it into a confusion matrix to better exploit the rich information inherent in hyperspectral images. Nevertheless, only using the spatial changes in images does not meet the requirements of practical urban planning and management applications. Therefore, in 2019, Liu et al. proposed a detection network for identifying spatio-temporal changes in urban slums [31], focusing not only on spatial changes but also on the dynamics of temporal changes. Although achieving good detection results, similar supervised methods typically require a large number of labeled samples for learning. Subsequently, this issue was addressed by Peng et al. with an unsupervised change detection method that utilizes visual saliency information to extract significant change regions [32]. This method reduces the influence of irrelevant information and enhances the representation ability of the change region. Additionally, the unsupervised detection framework significantly reduces the cost of manual labeling.

Recently, dual-branch networks have been developed in order to enhance the feature representation ability of deep learning models. In 2021, an asymmetric dual-branch network was introduced by Yang et al.; it utilizes features from different modules to accurately locate and identify semantic changes [33], which enhances the model's sensitivity in obtaining semantic information and improves change detection accuracy. Similarly, Zheng et al. [34] devised a cross-layer feature fusion module to combine feature maps from different branches, fully capturing the feature representation of change regions. However, these methods have notable drawbacks, such as poor feature fusion results and a lack of global information. The poor results of feature fusion prevent the model from accurately understanding the spatial relationships between pixels, which makes it difficult to determine the change regions and distinguish the boundary of the change areas. Consequently, the resulting map may be susceptible to false detection, missed detection, and blurred edge delineation of the change regions.

To address these issues, the proposed method incorporates feature interaction and edge constraints. On one hand, dual-temporal image feature interaction allows for capturing the context information between image pairs, and on the other hand, it can align the data distributions between the dual-temporal images, achieving data distribution adaptation. Further, the inclusion of edge constraints enables the model to better discern boundary information of the change area and reduce edge blurring.

2.2. Transformer-Based Change Detection Methods

In change detection, accurately identifying changes in dual-temporal images within a spatial and temporal range is crucial. It is essential to have a larger receptive field for capturing scene information and directing the model's attention towards the change regions with attention mechanisms. The CNN-based model proposed by Qian et al. [35] comprises stacked convolutional layers, extended convolutions, and attention mechanisms. Although this approach effectively captures local details, it struggles to connect distant details due to the limited receptive fields. Change detection methods based on Transformers can compensate for this shortcoming.

Currently, Transformer is widely employed across various image tasks. Unlike the fixed-size receptive field of convolutional neural networks, Transformer leverages its unique self-attention mechanism to gather global information, offering superior feature mapping and context modeling capabilities. In 2021, the BIT network was proposed [36], in which Transformer tokens are used to extract semantic information; then, a twin decoder structure is employed to highlight differences between two high-dimensional semantic tokens. Additionally, a multi-scale feature fusion module combines features from various scales and leverages a self-attention mechanism to weigh the fused features, enhancing sensitivity to the features of different layers. However, single-branch model structures have a limited perception of change features in dual-temporal images. In 2022, Bandara et al. proposed ChangeFormer [37], which is a two-branch model with a Transformer model that independently learns image features and better identifies change areas. Subsequently, the

VcT model was presented in 2023 [38]; it includes a new remote sensing change detection framework based on graph neural networks and Transformer. This model detects changes by dynamically adjusting attention to different regions based on an adaptive feature attention mechanism.

Although the Transformer model is strong at learning features and provides ample global information, which aids in effectively pinpointing change areas, its difficulty in accessing local details makes it challenging to accurately capture the boundaries of local changes. Furthermore, it remains a challenge to effectively learn meaningful feature representations of changes in remote sensing images in practical applications. To enhance the model's feature extraction ability, in this study, we established a dual-branch structure combining a CNN and Transformer. The CNN extracts local feature information, whereas Transformer can capture global information associations based on self-attention, extracting more global information. By leveraging this dual-branch structure, both the local and global information, extracted by the CNN and Transformer, respectively, can be simultaneously utilized for improving feature representation.

3. Methods

In this section, we provide a detailed description of the proposed model's specific structure. To address the issues of insufficient feature fusion and blurred boundaries in current change detection methods, we propose a model that employs cross-layer feature fusion and feature exchange. Our approach is based on the concept of feature interaction and integrates both the local information extracted by a CNN and the global information extracted by Transformer. The edge information obtained by the Sobel operator constrains the boundaries of the change area, enhancing detection accuracy and reducing edge blurring. Specific technical details and the technical roadmap are illustrated in Figure 1.

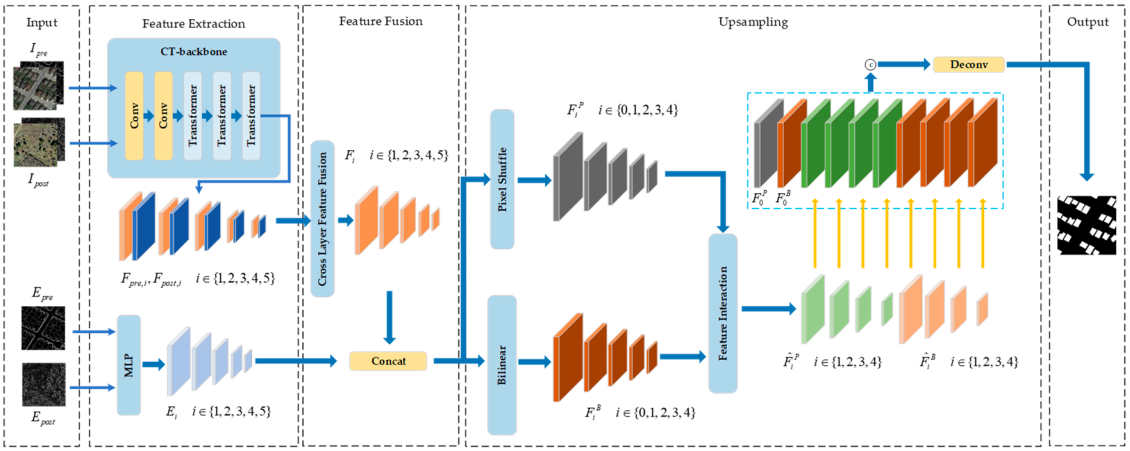

Figure 1. The overall structure diagram of the proposed model.

The proposed method encompasses feature extraction, feature fusion, and upsampling stages, similarly to the fundamental structure of current change detection methods. For the input dual-temporal images, the Sobel operator is first used to extract the edge information (E_{pre} and E_{post}) from the images. Subsequently, the edges are connected with the original image on the channel as the initial input, represented by I_{pre} and I_{post}, respectively.

After a dual-branch backbone network (CT-backbone), multiple scale feature maps are extracted, denoted by $F_{pre,i}$ and $F_{post,i}$, respectively, where $i \in \{1,2,3,4,5\}$ represents the size of the feature map relative to the original image. The specific correspondence is shown in Table 1. In the following, the same labels have the same meaning.

Table 1. Explanation of correspondence between labels and image sizes.

Original Picture	$i = 0$	$i = 1$	$i = 2$
$H \times W$	$H \times W$	$\frac{1}{2}H \times \frac{1}{2}W$	$\frac{1}{4}H \times \frac{1}{4}W$
	$i = 3$	$i = 4$	$i = 5$
	$\frac{1}{8}H \times \frac{1}{8}W$	$\frac{1}{16}H \times \frac{1}{16}W$	$\frac{1}{32}H \times \frac{1}{32}W$

Then, the extracted multi-scale feature information ($F_{pre,i}$ and $F_{post,i}$) is transmitted to the cross-layer feature information extraction module for fusion, aiming to enhance the representation of multi-scale features and attain an improved feature representation, denoted by F_i, with $i \in \{1, 2, 3, 4, 5\}$. Simultaneously, the edge information extracted by the Sobel operator, E_{pre} and E_{post}, is integrated through an MLP module to obtain E_i, with $i \in \{1, 2, 3, 4, 5\}$, as the edge constraint. Then, F_i and E_i are concatenated on channels and passed through a convolutional layer to extract the feature map, F'_i, with $i \in \{1, 2, 3, 4, 5\}$, which contains rich semantic and edge information. To effectively utilize both spatial and channel information, the proposed model simultaneously employs Bilinear and Pixel Shuffle methods for upsampling based on F'_i. F_i^B and F_i^P represent the results of the Bilinear and Pixel Shuffle methods, respectively, where $i \in \{0, 1, 2, 3, 4\}$. Subsequently, both of them are fed to the feature interaction module to fully integrate spatial and channel information. The outputs are expressed as $\hat{F}_i^P$ and $\hat{F}_i^B$, respectively, where $i \in \{0, 1, 2, 3, 4\}$. Finally, the outputs of feature fusion at different scales are upsampled to the same size at different magnifications. The final change map, F_{out}, is obtained by using a deconvolutional layer. The detailed process is as follows:

$$F_{out} = Deconv(Concat[UP(\hat{F}_i^P), UP(\hat{F}_i^B)]) \quad (1)$$

where F_{out} represents the model output and $UP(\cdot)$ represents the upsampling method of bilinear interpolation. The above is the main process of our model. Next, we will discuss the specific implementation details of each module in detail.

3.1. Dual-Branch Backbone Network Based on CNN and Transformer

To enhance the model's ability to extract image features, a dual-branch backbone using both a CNN and Transformer is proposed. The detailed structure is depicted in Figure 2.

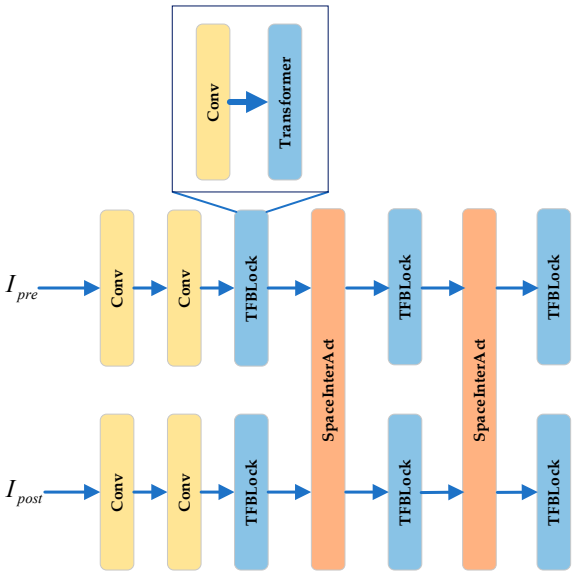

Figure 2. Main structure of backbone network based on CNN and Transformer.

For the input images I_{pre} and I_{post}, the proposed method uses two convolutional layers to extract shallow feature information:

$$F_{i,j+1} = ConvBlock_{s=2}^{k=3}(F_{i,j}),\ i \in \{pre, post\},\ j \in \{0, 1\} \tag{2}$$

where $ConvBlock_{s=2}^{k=3}$ represents a convolutional module with a kernel of 3×3, a stride of 2, and a padding of 1, and $F_{i,j}$ represents the feature map extracted in the jth stage of the pre-change image and the post-change image, where $F_{pre,0} = I_{pre}$ and $F_{post,0} = I_{post}$. Afterwards, Transformer is used to obtain deeper features.

$$F'_{i,j+1} = TFBlock(F_{i,j}),\ i \in \{pre, post\},\ j \in \{2, 3, 4\} \tag{3}$$

where $TFBlock$ consists of a downsampling module and a multi-head attention module in the Transformer model. The structure is as follows:

$$TFBlock(\cdot) = MHA(ConvBlock_{s=2}^{k=3}(\cdot)) \tag{4}$$

where $MHA(\cdot)$ represents the standard multi-head attention module [39]. Transformer is used to extract deep feature information; in order to further improve the feature extraction capabilities, a spatial feature interaction module is added.

Fang et al. [40] pointed out that feature interaction is essential to change detection. The core of change detection lies in detecting change areas with the same spatial position but different temporal characteristics. Distinguishing whether an image represents a scene before or after a change is merely useful for scholars to ensure that the disappearance and appearance of targets is reasonable. Whether those targets represented appear or disappear is meaningless to the model, for which change is the key. In other words, feature interaction does not revise the semantic information of change, making it feasible for change detection. On one hand, the model can perceive contextual information between image pairs by exchanging features. On the other hand, the data distributions between the dual-temporal images become more similar after feature exchange, and automatic adaptation of dual-temporal data distributions can be achieved. The specific process is illustrated as follows:

$$F_{pre,j+1}, F_{post,j+1} = SpaceInterAct(F'_{pre,j}, F'_{post,j}),\ j \in \{2, 3, 4\} \tag{5}$$

$SpaceInterAct(\cdot)$ represents spatial information exchange. The specific implementation is as follows:

$$x_{pre/post}(n, c, h, w) = \begin{cases} x_{pre/post}(n, c, h, w), M(n, c, h, w) = 0 \\ x_{post/pre}(n, c, h, w), M(n, c, h, w) = 1 \end{cases} \tag{6}$$

where n, c, and w denote the batch size, channel number, and spatial size, respectively. M represents an exchange mask composed of 1 and 0, indicating areas for exchange and non-exchange. In this model, the weight map output from the multi-head attention module is the criterion for determining whether an exchange should occur. If the weight in the output map exceeds threshold δ, it is a region to be exchanged, setting $M = 1$; otherwise, $M = 0$.

By using feature interaction, the dual-branch backbone network based on the CNN and Transformer can extract both local and global information while enhancing interaction among features through spatial information exchange. This leads to a greater similarity in the data distributions between pre- and post-change images, thereby further highlighting the change areas.

3.2. Cross-Layer Feature Fusion

In deep learning, features at different levels often correspond to spatial information at different scales. With cross-layer feature fusion, it is possible to effectively integrate

features at different scales, providing the model with a richer and more diverse feature representation, thereby aiding the model to more comprehensively understand the semantic information in the images. After the backbone, the model extracts feature maps $F_{pre,i}$ and $F_{post,i}$ at various scales (where $i \in \{1,2,3,4,5\}$) from the dual-temporal images. Subsequently, these feature maps are directed to the cross-layer feature fusion module for comprehensive feature fusion, which is illustrated in Figure 3.

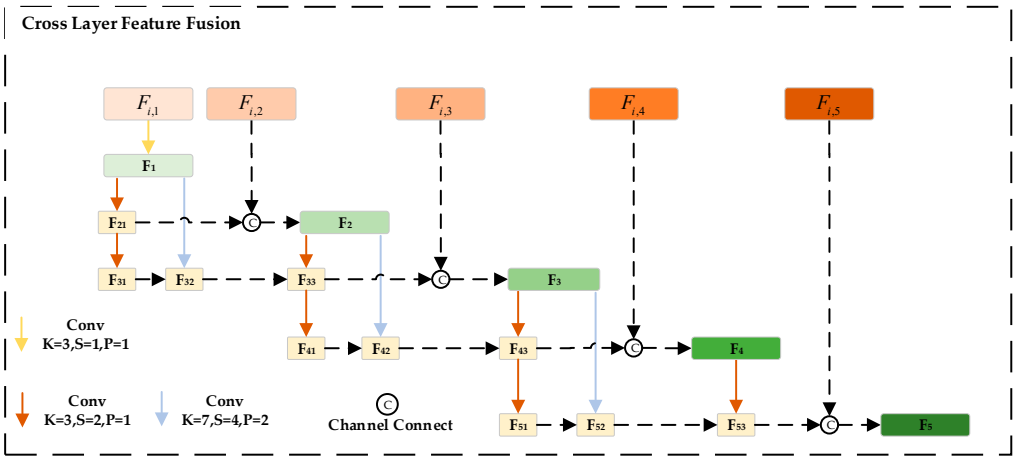

Figure 3. Cross-layer feature fusion module.

In Figure 3, feature map $F_{i,1}$ is subjected to a convolution operation to extract more abstract feature representations, expressed as F_1. Subsequently, the multiple-scale feature maps F_{21}, F_{31}, F_{32} are extracted with two branches. The left branch (indicated by the orange arrow) involves downsampling the feature map twice, while the right branch (indicated by the light-blue arrow) involves downsampling the feature map four times. Afterwards, the obtained F_{21} and initial input $F_{i,1}$ are concatenated along the channels to obtain F_2, which integrates multi-scale features and multi-level contextual information. The same method is subsequently applied to F_3, F_4, F_5.

The dual-temporal images $F_{pre,i}$ and $F_{post,i}$ undergo comprehensive feature fusion in the cross-layer feature fusion module, yielding $\hat{F}_{pre,i}$ and $\hat{F}_{post,i}$, where $i \in \{1,2,3,4,5\}$. Then, $\hat{F}_{pre,i}$ and $\hat{F}_{post,i}$ are concatenated along the channel dimension and passed through a convolutional module to generate F_i, which integrates multi-scale feature information. The specific process is outlined as follows:

$$F_i = ConvBlock(Concat[\hat{F}_{pre,i}, \hat{F}_{post,i}]),\ i \in \{1,2,3,4,5\} \tag{7}$$

In contrast to CLNet, the cross-layer fusion method accomplishes the extraction and fusion of multi-scale features through the utilization of two asymmetric branches. This approach ensures that the intermediate feature map captures both higher-level and lower-level context information, enhancing the model's feature-capturing capabilities. The improved extraction and representation aid the model in effectively discerning change areas.

3.3. MLP-Based Edge Information Extraction Module

To enhance accuracy in predicting change area boundaries, an edge information extraction module based on the MLP structure was incorporated. MLP includes an input layer, a hidden layer, and an output layer. In the hidden layer, neurons from the preceding layer are fully connected to neurons in the subsequent layer, forming a comprehensive fully connected structure [41]. This full connection facilitates the aggregation of features to a significant extent. In the experiment, the number of hidden layers was set to 2 for the preliminary aggregation of fragmented edges. Simultaneously, the GELU activation function was employed to accelerate the convergence of the model.

Due to the lack of semantic information, the simple edge information (E_{pre} and E_{post}) from the dual-temporal images obtained by using the Sobel operator often results fragmented and not connected. To solve this, the MLP module is employed to amalgamate features from fragmented edge information, yielding edge details across multiple scales. The hidden layers of the MLP map input features through nonlinear activation functions, allowing complex nonlinear transformations in multidimensional space. This enables the MLP to capture richer and more complex patterns and information from input features. Additionally, each neuron in every hidden layer of the MLP is connected to all neurons in the previous layer, with each connection having a weight parameter. These connectivity and weight parameters facilitate information propagation and feature aggregation between different hidden layers, enabling the MLP to effectively extract rich information from input features and aggregate it into higher-level representations throughout the network. This strategy enables detailed edge information capture in change areas and mitigates edge blurring. The process is detailed below:

$$E_{i+1} = MLP(ConvBolck_{s=2}^{k=3}(E_i)) \tag{8}$$

where E_i represents the output of the i-th layer of the MLP structure. Finally, five hierarchical edge features $[E_1, E_2, E_3, E_4, E_5]$ are extracted. Subsequently, these features are fused with features F_i obtained from the cross-layer feature fusion stage and inputted into the upsampling module for upsampling.

3.4. Upsampling and Prediction Module

After the feature extraction and fusion module, feature maps F_i' are obtained by merging various levels of edge information and multi-scale feature maps, where $i \in \{1, 2, 3, 4, 5\}$. To improve spatial and channel information integration, the Pixel Shuffle and Bilinear upsampling methods are utilized for the individual upsampling of the feature maps. The detailed process is outlined as follows:

$$\begin{aligned} F_i^P &= PixelShuffle(F_i) \\ F_i^B &= Bilinear(F_i) \end{aligned} \tag{9}$$

where F_i^P and F_i^B denote the outcomes of the upsampling using the Pixel Shuffle and Bilinear methods, where $i \in \{0, 1, 2, 3, 4\}$. Then, the channel information interaction module is applied to exchange channel information on the upsampled results, F_i^P and F_i^B, as illustrated in Figure 4 in detail.

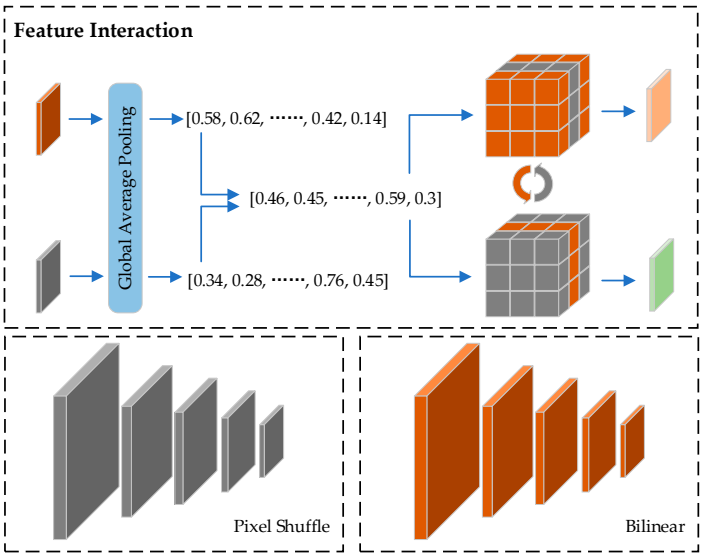

Figure 4. Channel feature interaction module.

For instance, for the feature maps F_i^P and F_i^B of the i-th layer (with size $h_i \times w_i \times c_i$), the initial step involves transforming them into $1 \times 1 \times c_i$ by using global average pooling. Subsequently, applying softmax converts them into weights, selecting channels whose average values exceeds η for channel exchange, resulting in the exchanged feature maps $\hat{F}_i^P$ and $\hat{F}_i^B$. Through channel information exchange, the model can be promoted to capture the information of each feature more comprehensively. Meanwhile, by improving the diversity and richness of features, the final features are made more distinguishable. Afterwards, Bilinear is applied 16, 8, 4, and 2 times to restore them to the original image size. Finally, the obtained multiple feature maps are concatenated along the channels, and the final output is obtained by using a deconvolution.

3.5. Loss Function

For change detection models, MSE loss is widely used, as it can effectively assess the performance of change detection models on the entire image. The proposed model also uses it as part of the loss function, which is defined as follows:

$$L_{mse} = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2 \tag{10}$$

As datasets may exhibit an imbalance between the number of positive and negative samples, we incorporated dice loss into the loss function. Dice loss is known for its sensitivity to imbalanced data, which serves to mitigate the effects of data imbalance, and is defined as follows:

$$L_{dice} = 1 - \frac{2 \times (\sum_{i=1}^{n} y_i \times p_i + smooth)}{\sum_{i=1}^{n} y_i + p_i + smooth} \tag{11}$$

In the above equation, n represents the total number of pixels; y_i and p_i represent the real change map and the model prediction map, respectively, and their values range from 0 to 1. To prevent the occurrence of non-change regions in dice loss, a parameter smoothing factor is considered, where $smooth = 1$. The final loss is, therefore, defined as follows:

$$L = L_{mse} + L_{dice} \tag{12}$$

4. Results

4.1. Datasets and Experimental Setup

The experiments were conducted with three public datasets: LEVIR-CD [42], WHU Building [43], and xBD [44].

The LEVIR-CD dataset collects Google Earth remote sensing images of multiple cities, including Austin and Lakeway in Texas, USA. It presents a large number of illumination changes due to seasonal effects, and the building change areas are small and dense, which makes it more challenging to determine the actual change areas. The WHU Building dataset was proposed by the Wuhan University team. Compared with the LEVIR-CD dataset, it has larger buildings, and the change areas are sparser. The xBD dataset was proposed by MIT and contains remote sensing images before and after 19 natural disasters such as earthquakes, volcanoes, and floods, and the change areas are mostly irregular, making detection more difficult.

In our experiments, the original images were cropped into non-overlapping 256×256 sections and then randomly allocated to training, validation, and test sets in a ratio of 7:2:1 for experimentation. Notably, the xBD dataset classifies change areas into four damage levels: no damage, minor damage, major damage, and destroyed. As our focus lies solely in change areas, we treated the latter three classes as change during experimentation. The models were trained from scratch for 30 epochs, using an initial learning rate of 0.001 and a batch size of 16. The learning rate decreased by 10% every 5 iterations after the

initial 15 iterations. The hyperparameters for spatial feature exchange and channel feature exchange were set to 0.5. To ensure an equitable performance comparison, the loss functions of all the comparative methods were replaced with the proposed model's loss function, neutralizing performance discrepancies due to varied loss functions.

4.2. Comparison Method

DSIFN [45]: A change detection model that uses cross-layer connections for feature fusion.

SNUNet [46]: A change detection model employing multi-layer feature fusion with dense connections that combines a Siamese network and NestedUNet to extract sophisticated features and incorporates channel attention and deep supervision techniques to enhance the recognition ability of intermediate features.

BIT [36]: A detection model based on Transformer that uses Transformer to build an encoder–decoder structure, enhances the feature information of the context through semantic tokens and feature differences, and obtains the change map.

ChangeFormer [38]: A change detection model that enhances its feature extraction capabilities by replacing the convolutional neural network with Transformer. Additionally, it utilizes the MLP structure to enhance feature differences.

SGSLN [47]: A novel strategy involving the swapping of dual codec backbones for binary change detection. A temporal fusion attention module is employed to effectively fuse dual-temporal features for enhanced detection.

4.3. Evaluation Metrics

To quantitatively assess the models' performance, five evaluation metrics were selected to measure the disparities between the predicted change maps and the actual change maps. The chosen indicators included precision (P), recall (R), F1 score ($F1$), overall accuracy (OA), and average intersection over union ($mIoU$). Precision (P) is the ratio of correctly predicted changed pixels to all predicted changed pixels, while recall (R) denotes the ratio of the overall true changed pixels. F1 score is the harmonic mean of precision and recall. Overall accuracy (OA) reflects the proportion of correctly predicted pixels to the entire pixel count. The average intersection over union ($mIoU$) provides a comprehensive assessment of detection performance for both change and non-change areas. The calculation equations for these five indicators are as follows:

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = \frac{2 \times TP \times TN}{TP + FN} \tag{15}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

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

In the above, TP is the true value, TN is the true negative value, FP is the false positive value, and FN is the false negative value.

4.4. Results and Discussion

The experimental results show the adaptability of the proposed model across diverse change scenarios of varying scales. It not only demonstrated impressive performance on both the WHU Building and xBD datasets, which have larger structures, but also on the xBD dataset, which has smaller change areas. The results are here discussed in terms of both quantitative and qualitative aspects.

Quantitative results: As shown in Tables 2 and 3, the proposed model outperformed the other five models, achieving optimal scores across the five evaluation indicators. On the LEVIR and xBD datasets, while the precision (P), recall (R), and F1 scores of the proposed model show only marginal improvement over the second-best method, the mIoU index exhibits a notable increase of approximately 1.5 percent compared with other models. This is attributed to the inclusion of the boundary constraint module, which heightens the model's sensitivity to change area edges through boundary constraints. Consequently, the blurring of the edges and the connection of the areas are reduced, aligning the predicted change areas more closely with their actual shapes. The change areas in the WHU dataset are larger, and their edges exhibit more regular shapes, so the proposed model, on the WHU dataset, outperformed the other methods, which have no explicit edge constraints, in terms of accuracy and mIoU.

Table 2. Quantitative comparison results of the model on the LEVIR and WHU Building datasets. The highest index is shown in bold, and the second-highest index is underlined.

Method	LEVIR					WHU				
	P	R	F1	OA	mIoU	P	R	F1	OA	mIoU
SNUNet	0.9274	0.9083	0.9170	99.28	0.8479	0.8828	0.8705	0.8694	99.17	0.7946
DSIFN	0.9303	0.9147	0.9218	99.33	0.8564	0.9023	0.9088	0.8989	99.38	0.8378
BIT	0.9234	0.9094	0.9162	99.25	0.8456	0.8862	0.8683	0.8694	99.15	0.7936
ChangeFormer	0.9195	0.9009	0.9095	99.23	0.8354	0.8131	0.8010	0.7899	98.62	0.6872
SGSLN	0.9200	0.9038	0.9116	99.21	0.8379	0.8828	0.8886	0.8757	99.12	0.7982
Ours	0.9377	0.9198	0.9279	99.38	0.8669	0.9231	0.9150	0.9149	99.53	0.8651

Table 3. Quantitative comparison results of the model on the xBD disaster dataset, with the highest indicator in bold and the second-highest indicator underlined.

Method	xBD				
	P	R	F1	OA	mIoU
SNUNet	0.9226	0.9206	0.9216	94.65	0.8547
DSIFN	0.9334	0.9331	0.9332	95.43	0.8618
BIT	0.9005	0.9018	0.9010	93.23	0.8201
ChangeFormer	0.9144	0.9118	0.9130	94.05	0.8400
SGSLN	0.8928	0.8927	0.8926	92.65	0.8062
Ours	0.9336	0.9343	0.9345	95.49	0.8771

Qualitative results: As shown in Figure 5, on the LEVIR dataset, the change areas in images a and b exhibit denser and more regular boundaries. In image a, the SGSLN model exhibits suboptimal detection in the region highlighted by the blue box. This is due to the influence of the house shadow, resulting in fragmented results and an inability to accurately delineate the change area. Similarly, the SNUNet, DSIFN, BIT, and ChangeFormer methods are also affected by the shadow in this region, exhibiting varying degrees of overlap in their detection outcomes and poorly distinguishing change areas. In contrast, the proposed method demonstrated superior visual performance with minimal connected areas. Similarly, in image b, the proposed model outperformed the others significantly in the yellow box. The change area in image c presents an irregular shape, posing greater detection challenges than the first two images. However, the results illustrate that the proposed model exceled at capturing the region and preserving the shape of the change area, exhibiting no instances of missed detection or blurred boundaries. The WHU Building dataset features larger change areas with a more regular pattern than the LEVIR dataset. Figure 6 shows the results for the WHU dataset. In images e and f, it is evident that the proposed model provided more comprehensive predictions of change areas and achieved greater accuracy at the boundaries.

The xBD dataset presents denser and smaller change areas characterized by numerous irregular shapes than the LEVIR dataset. Similar to the above, superior results were achieved by the proposed model when facing these challenges. As shown in Figure 7, the region highlighted by the green box in image g reveals that all the models except ChangeFormer generated false detection results, erroneously identifying the top portion of land as a change area. Despite ChangeFormer having better performance in discerning change areas, its edge prediction notably lagged behind that of the proposed method. Likewise, within the green-marked area in image h, only the proposed method achieved exceptional detection outcomes for the irregular segments within the change area.

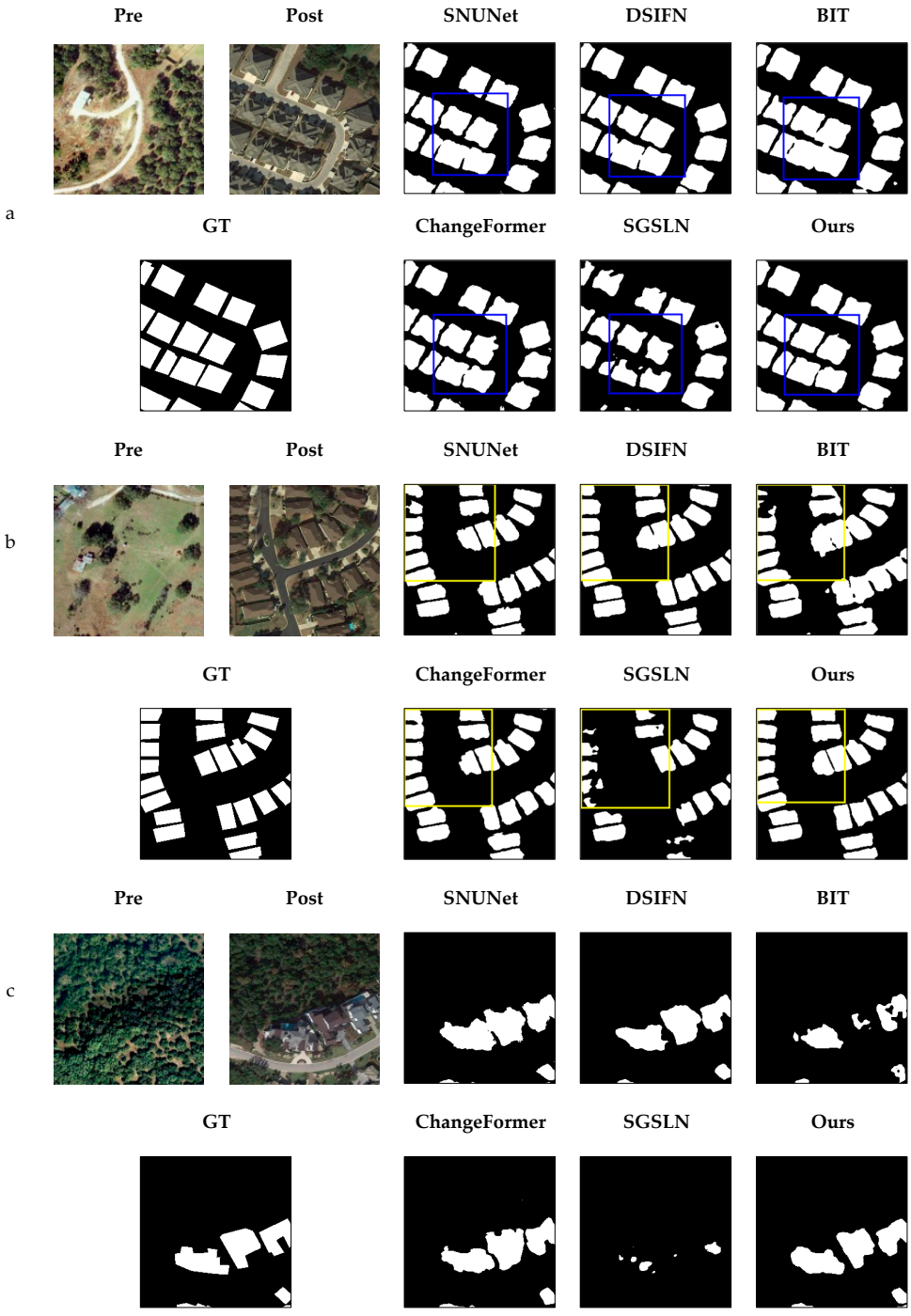

Figure 5. Visualization results of different methods on LEVIR dataset.

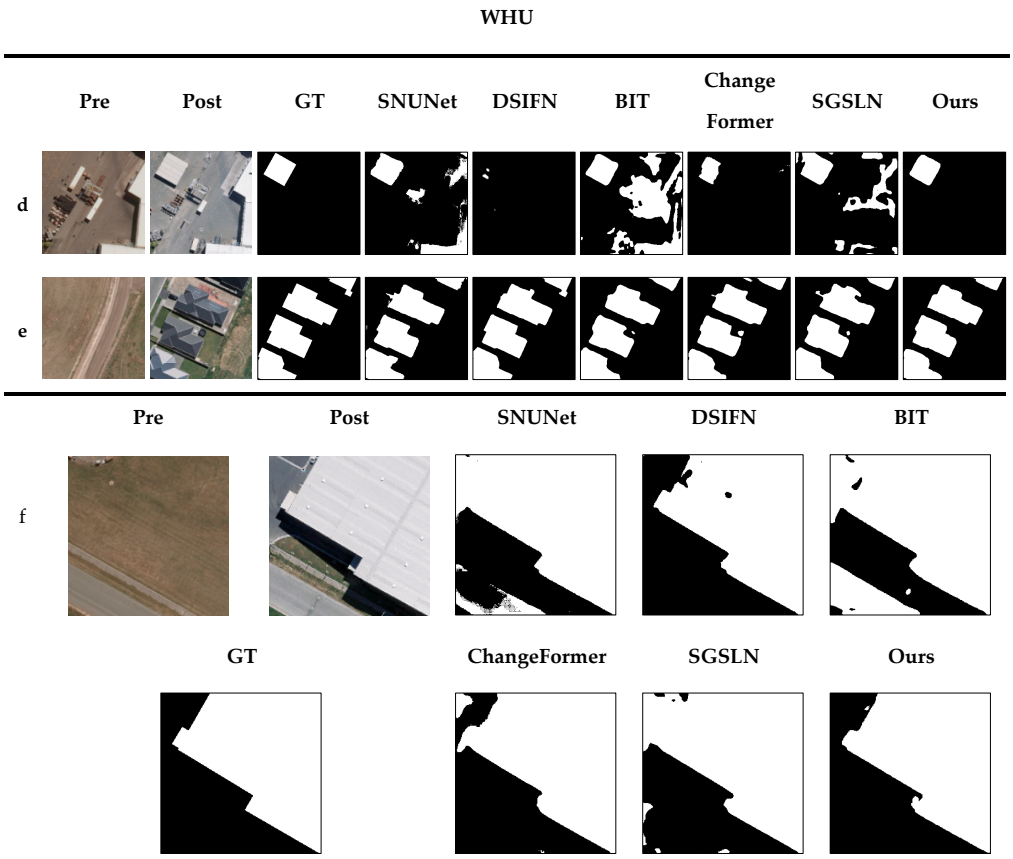

Figure 6. Visualization results of different methods on WHU dataset.

Overall, whether it was the xBD dataset with small and dense change areas, the LEVIR dataset with more common change area shapes, or the WHU dataset with larger change areas, the proposed model achieved superior outcomes. This is largely attributed to the efficacy of our boundary constraint module. By integrating boundary constraints, the proposed model achieves two key objectives: On one hand, it effectively discriminates among various change areas in dense regions and reduces regional overlap. On the other hand, it ensures that the predicted boundaries closely match their actual values.

To further validate the effectiveness of the model, distinct colors were employed to represent true positive (TP; white), true negative (TN; black), false positive (FP; red), and false negative (FN; green) results, as depicted in Figure 8. The proposed model outperformed the others in various aspects. It effectively avoided false positive (FP) instances, indicated by the red regions. In images a, d, and h, the proposed model closely approximated the real values along the edges. Moreover, it significantly reduced the occurrence of missed detection, evident in the fewer green regions compared with the results of the other methods. This distinction is particularly noticeable in images a, c, and i.

To evaluate the effectiveness of the edge constraint module, ablation experiments were conducted on the LEVIR dataset. Table 4 showcases the quantitative findings, and Figure 9 illustrates the qualitative results. The model with the edge constraint module demonstrated improvements across different metrics compared with the model without it, with a notable increase in the mIoU metric. This improvement highlights the role of the edge constraint module in accurately predicting the change region. Visually, the change areas closely aligned with the true values at the edges, providing empirical evidence of the efficacy of the boundary constraint module.

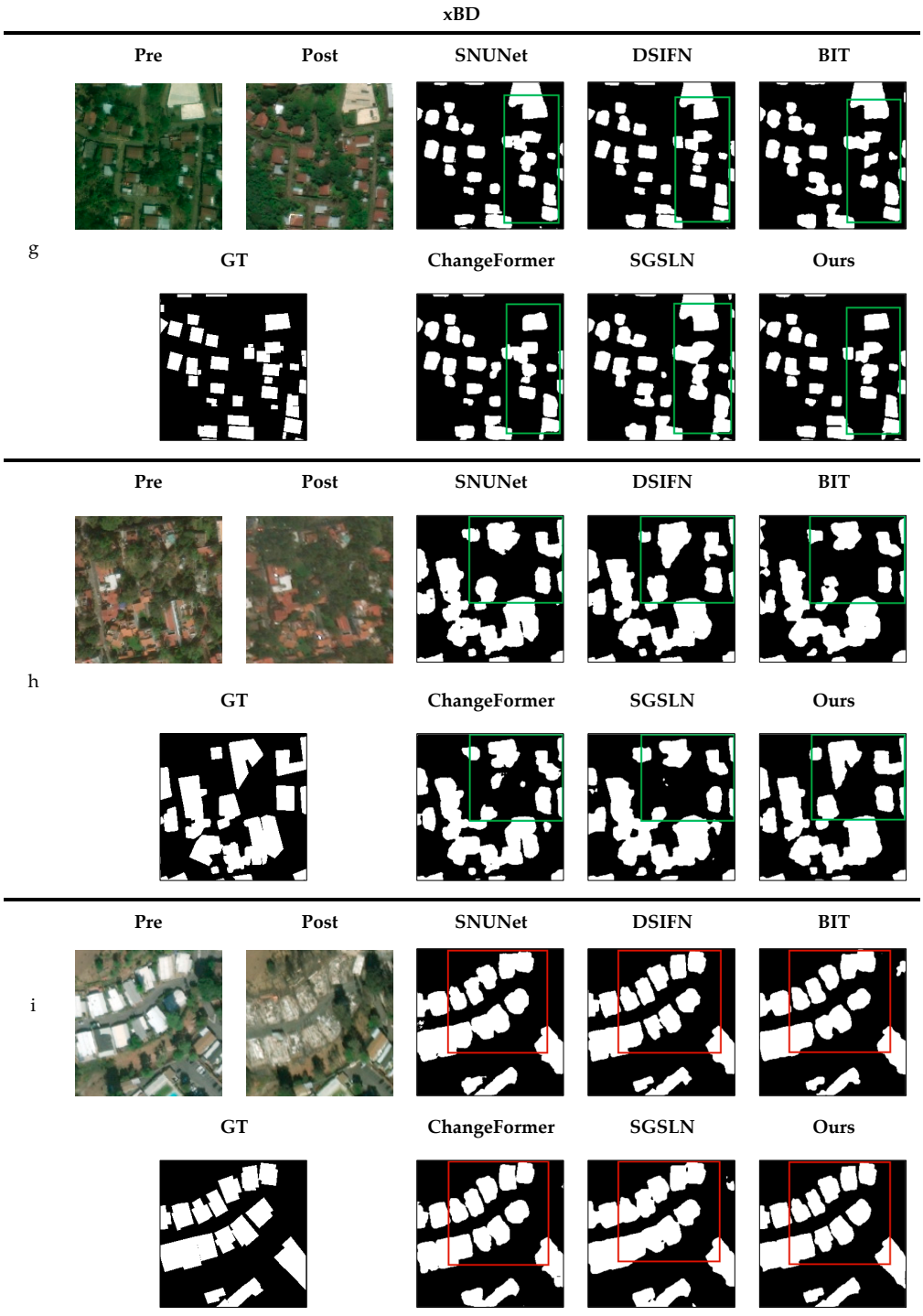

Figure 7. Visualization results of different methods on xBD datasets.

Furthermore, the influence of employing two upsampling methods during the upsampling stage was taken into account. According to the results in Table 4, it is clear that the combined use of the Pixel Shuffle and Bilinear upsampling methods can significantly boost detection accuracy. This increase stems from the concurrent integration of channel and spatial information, thereby improving the model's capability to precisely capture change regions.

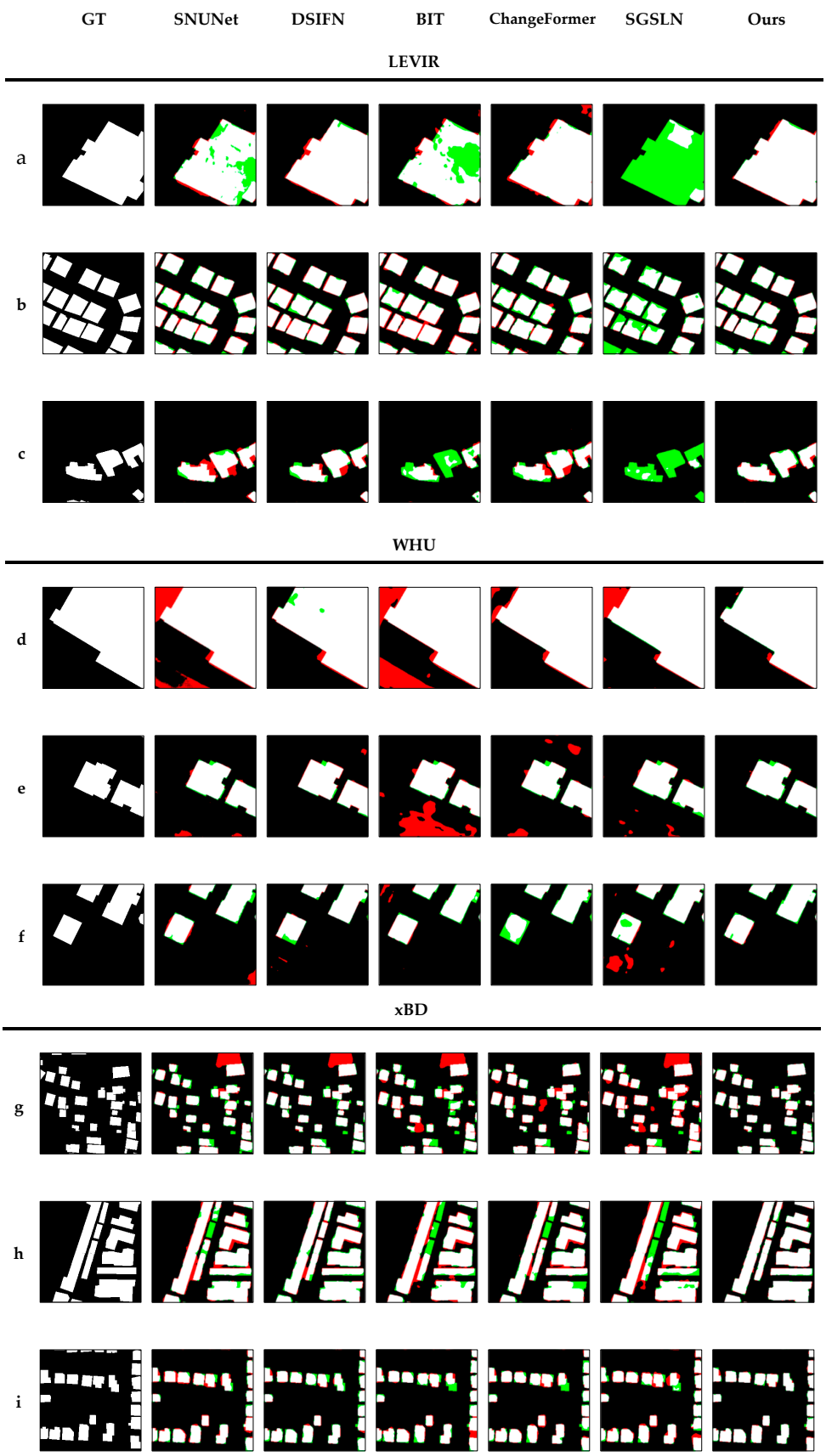

Figure 8. Visualization results of different methods on LEVIR, WHU Building, and xBD datasets, including TP (white), TN (black), FP (red), and FN (green).

Table 4. Edge constraint module ablation experimental results.

Edge	Pixel Shuffle	Bilinear	P	R	F1	mIoU
	√	√	0.9324	0.9149	0.9230	0.8581
√		√	0.9366	0.9181	0.9266	0.8644
√	√		0.9348	0.9187	0.9260	0.8645
√	√	√	0.9377	0.9198	0.9279	0.8669

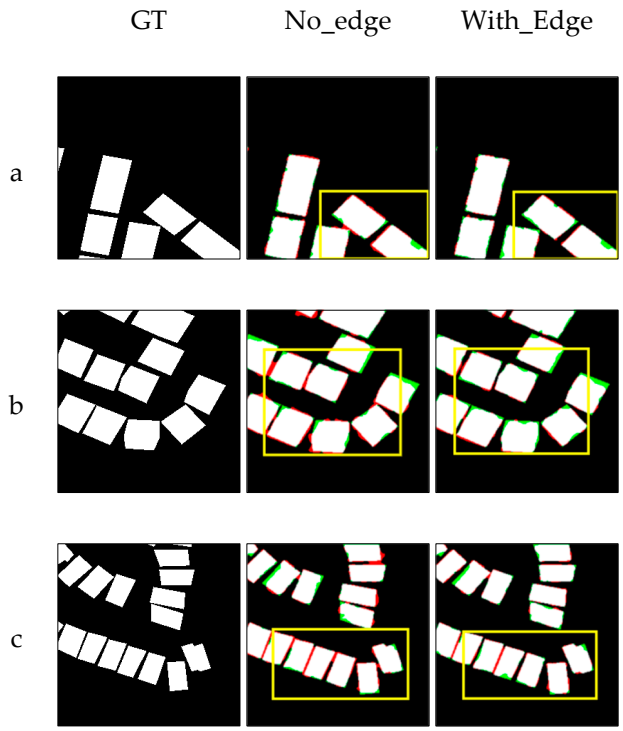

Figure 9. Comparison among qualitative results highlighting the contribution of the edge constraint module, including TP (white), TN (black), FP (red), and FN (green).

5. Discussion

Selecting accurate models and algorithms is crucial for change detection. By applying precise algorithms or models in change detection, the detection accuracy can be improved, achieving more desirable results. This aligns with the current development trends in change detection.

Although previous studies based on CNN networks have demonstrated the powerful feature extraction capabilities of deep learning methods, there are still issues with the clarity of boundaries in the identified regions. This shortcoming is mainly due to the inadequate aggregation of contextual information during feature extraction. The network proposed in this study achieves contextual feature aggregation through the use of a cross-layer feature fusion module and significantly enhances the precision of change regions by integrating spatial and channel information via a dual-branch upsampling module. Additionally, the introduction of a boundary constraint module, which consolidates fragmented edge information through an MLP module, effectively increases boundary constraints within change regions and reduces boundary blurring. These improvements are not only academically significant but also provide more precise and reliable solutions for practical change detection tasks, especially in natural disaster assessment and urban building change monitoring.

Despite the superiority of our method across multiple datasets, there is still room for further improvement. Future research can be carried out in the following aspects: First, more types of datasets and application scenarios can be explored to verify the generality and adaptability of the method. Second, although Transformers effectively capture global information, their computational requirements pose challenges for model training. To

advance and extend the proposed method, future research could explore lightweight change detection methods aimed at enhancing the practicality and efficiency of existing methods.

6. Conclusions

Currently, change detection is a central focus in the field of remote sensing. To address feature fusion deficiency and the challenge of blurred edges in the detected change areas, a dual-branch change detection model was introduced. In the feature extraction stage, it combines the local information derived from the CNN with the global information from Transformer, and the feature information is fully fused through the cross-layer feature fusion module, solving the issue of the absence of comprehensive global information in existing feature fusion methodologies. Additionally, a boundary constraint module based on an MLP was introduced to process the edge information of the change areas, which mitigates edge blurring. Moreover, spatial and channel information were integrated to bolster detection accuracy by using two upsampling methods. The experimental findings across three datasets show the model's ability to achieve high precision while proficiently delineating the boundaries of the change areas.

Author Contributions: All the authors made significant contributions to the work. X.W. and Z.G. designed this research study, analyzed the results, and performed the validation work. R.F. provided advice for the revision of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the China Scholarship Council, the Hubei Key Laboratory of Intelligent Geo-Information Processing (No. KLIGIP-2019B08), the Sub-pixel Mapping of Hyperspectral Remote Sensing Images Based on Deep Unfolding Networks (2024AFB561), the Knowledge Innovation Program of Wuhan–Shuguang (202301020102336), and the National Natural Science Foundation of China under Grant (41925007).

Data Availability Statement: The data presented in this study are available in reference number [42–44].

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

References

1. Gao, Y.; Gao, F.; Dong, J.; Li, H. SAR image change detection based on multiscale capsule network. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 484–488. [CrossRef]
2. Alizadeh, N.A.; Beirami, B.; Mokhtarzade, M. Damage detection after the earthquake using Sentinel-1 and 2 images and machine learning algorithms (case study: Sarpol-e Zahab earthquake). In Proceedings of the 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 17–18 November 2022; pp. 343–347.
3. Wu, K.; Ma, Y.; Zhang, L. Sub-pixel land-cover change detection based on pixel unmixing and EM algorithm. In Proceedings of the 2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Tokyo, Japan, 2–5 June 2015; pp. 1–4.
4. Wang, L.; Zuo, B.; Le, Y.; Chen, Y.; Li, J. Penetrating remote sensing: Next-generation remote sensing for transparent earth. *Innovation* **2023**, *4*, 100519. [CrossRef] [PubMed]
5. Zhou, Y. Research on Forest resource change detection based on decision tree algorithm. In Proceedings of the 2022 International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS), Bristol, UK, 29–31 July 2022; pp. 363–367.
6. Zhang, W.; Fan, H. Application of isolated forest algorithm in deep learning change detection of high resolution remote sensing image. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 24–26 June 2022; pp. 753–756.
7. Shafique, A.; Cao, G.; Khan, Z.; Asad, M.; Aslam, M. Deep learning-based change detection in remote sensing images: A review. *Remote Sens.* **2022**, *14*, 871. [CrossRef]
8. Zhang, F.; Liu, K.; Liu, Y.; Wang, C.; Zhuo, W.; Zhang, H.; Wang, L. Multitarget Domain Adaptation Building Instance Extraction of Remote Sensing Imagery With Domain-Common Approximation Learning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4702916. [CrossRef]
9. Shi, W.; Zhang, M.; Zhang, R.; Chen, S.; Zhan, Z. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sens.* **2020**, *12*, 1688. [CrossRef]
10. Jiang, H.; Peng, M.; Zhong, Y.; Xie, H.; Hao, Z.; Lin, J.; Ma, X.; Hu, X. A Survey on deep learning-based change detection from high-resolution remote sensing images. *Remote Sens.* **2022**, *14*, 1552. [CrossRef]
11. Shen, L.; Lu, Y.; Chen, H.; Wei, H.; Xie, D.; Yue, J.; Chen, R.; Lv, S.; Jiang, B. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sens.* **2021**, *13*, 5094. [CrossRef]

12. Hao, M.; Zhang, H.; Shi, W. Unsupervised change detection using fuzzy c-means and MRF from remotely sensed images. *Remote Sens. Lett.* **2013**, *4*, 1185–1194. [CrossRef]
13. Lei, T.; Zhang, Y.; Lv, Z.; Li, S.; Liu, S.K.; Nandi, A. Landslide Inventory Mapping from Bitemporal Images Using Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 982–986. [CrossRef]
14. Lin, T.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
15. Zhang, C.; Wei, S.; Ji, S.; Lu, M. Detecting large-scale urban land cover changes from very high resolution remote sensing images using cnn-based classification. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 189. [CrossRef]
16. Wu, C.; Zhang, F.; Xia, J.; Xu, Y.; Li, G.; Xie, J.; Du, Z.; Liu, R. Building damage detection using u-net with attention mechanism from pre- and post-disaster remote sensing datasets. *Remote Sens.* **2021**, *13*, 905. [CrossRef]
17. Zhang, Y.; Fu, L.; Li, Y.; Zhang, Y. HDFNet: Hierarchical dynamic fusion network for change detection in optical aerial images. *Remote Sens.* **2021**, *13*, 1440. [CrossRef]
18. Hussain, M.; Chen, D.; Cheng, A.; Wei, H.; Stanley, D. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS J. Photogramm. Remote Sens.* **2013**, *80*, 91–106. [CrossRef]
19. Patra, R.K.; Patil, S.N.; Falkowski-Gilski, P.; Łubniewski, Z.; Poongodan, R. Feature weighted attention—Bidirectional long short term memory model for change detection in remote sensing images. *Remote Sens.* **2022**, *14*, 5402. [CrossRef]
20. 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]
21. Lin, Y.; Li, S.; Fang, L.; Chamisi, P. Multispectral change detection with bilinear convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1757–1761. [CrossRef]
22. Li, H.; Gong, M.; Zhang, M.; Wu, Y. Spatially self-paced convolutional networks for change detection in heterogeneous images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4966–4979. [CrossRef]
23. Wang, G.; Li, B.; Zhang, T.; Zhang, S. A Network combining a transformer and a convolutional neural network for remote sensing image change detection. *Remote Sens.* **2022**, *14*, 2228. [CrossRef]
24. Zhang, B.; Ye, H.; Lu, W.; Huang, W.; Wu, B.; Hao, Z.; Sun, H. A spatiotemporal change detection method for monitoring pine wilt disease in a complex landscape using high-resolution remote sensing imagery. *Remote Sens.* **2021**, *13*, 2083. [CrossRef]
25. Zhu, Y.; Tang, H. Automatic damage detection and diagnosis for hydraulic structures using drones and artificial intelligence techniques. *Remote Sens.* **2023**, *15*, 615. [CrossRef]
26. Zhan, T.; Song, B.; Xu, Y.; Wan, M.; Wang, X.; Yang, G.; Wu, Z. SSCNN-S: A spectral-spatial convolution neural network with siamese architecture for change detection. *Remote Sens.* **2021**, *13*, 895. [CrossRef]
27. Chen, D.; Wang, Y.; Shen, Z.; Liao, J.; Chen, J.; Sun, S. Long time-series mapping and change detection of coastal zone land use based on google earth engine and multi-source data fusion. *Remote Sens.* **2022**, *14*, 1. [CrossRef]
28. Zhang, H.; Wang, M.; Wang, F.; Yang, G.; Zhang, Y.; Jia, J.; Wang, S. A novel squeeze-and-excitation w-net for 2D and 3D building change detection with multi-source and multi-feature remote sensing data. *Remote Sens.* **2021**, *13*, 440. [CrossRef]
29. Mastro, P.; Masiello, G.; Serio, C.; Pepe, A. Change detection techniques with synthetic aperture radar images: Experiments with random forests and Sentinel-1 observations. *Remote Sens.* **2022**, *14*, 3323. [CrossRef]
30. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3–13. [CrossRef]
31. Liu, R.; Kuffer, M.; Persello, C. The temporal dynamics of slums employing a CNN-based change detection approach. *Remote Sens.* **2019**, *11*, 2844. [CrossRef]
32. Peng, D.; Guan, H. Unsupervised change detection method based on saliency analysis and convolutional neural network. *J. Appl. Remote Sens.* **2019**, *13*, 024512. [CrossRef]
33. Yang, K.; Xia, G.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; Zhang, L. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5609818. [CrossRef]
34. Zheng, Z.; Wan, Y.; Zhang, Y.; Xiang, S.; Peng, D.; Zhang, B. CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 247–267. [CrossRef]
35. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A Deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5604816. [CrossRef]
36. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
37. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection IGARSS 2022. In Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210.
38. Jiang, B.; Wang, Z.; Wang, X.; Zhang, Z.; Chen, L.; Wang, X.; Luo, B. VcT: Visual change transformer for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 2005214. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
40. Fang, S.; Li, K.; Li, K. Changer: Feature interaction is what you need for change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610111. [CrossRef]

41. Taud, H.; Mas, J.F. Multilayer perceptron (MLP). In *Geomatic Approaches for Modeling Land Change Scenarios*; Springer International Publishing: Cham, Switzerland, 2018; pp. 451–455.
42. 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]
43. 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]
44. Gupta, R.; Hosfelt, R.; Sajeev, S.; Patel, N.; Goodman, B.; Doshi, J.; Heim, E.; Choset, H.; Gaston, M. Creating xBD: A dataset for assessing building damage from satellite imagery. In Proceedings of the of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, California, CA, USA, 16–17 June 2019.
45. 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]
46. 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*, 8007805. [CrossRef]
47. Zhao, S.; Zhang, X.; Xiao, P.; He, G. Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4508016. [CrossRef]