remote sensing

MDPI

Article

MSGFNet: Multi-Scale Gated Fusion Network for Remote Sensing Image Change Detection

Yukun Wang [1], Mengmeng Wang [2,*], Zhonghu Hao [1], Qiang Wang [1], Qianwen Wang [3] and Yuanxin Ye [2]

[1] School of Mechatronics Engineering, Beijing Institute of Technology, Beijing 100081, China; 3120185169@bit.edu.cn (Y.W.); haozhonghu@bit.edu.cn (Z.H.); wang_qiang@bit.edu.cn (Q.W.)
[2] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China; yeyuanxin@home.swjtu.edu.cn
[3] The Fifth Military Delegate Office in Beijing, Beijing 100038, China; dfhxdg1@163.com
* Correspondence: wm_email@my.swjtu.edu.cn

Abstract: Change detection (CD) stands out as a pivotal yet challenging task in the interpretation of remote sensing images. Significant developments have been witnessed, particularly with the rapid advancements in deep learning techniques. Nevertheless, challenges such as incomplete detection targets and unsmooth boundaries remain as most CD methods suffer from ineffective feature fusion. Therefore, this paper presents a multi-scale gated fusion network (MSGFNet) to improve the accuracy of CD results. To effectively extract bi-temporal features, the EfficientNetB4 model based on a Siamese network is employed. Subsequently, we propose a multi-scale gated fusion module (MSGFM) that comprises a multi-scale progressive fusion (MSPF) unit and a gated weight adaptive fusion (GWAF) unit, aimed at fusing bi-temporal multi-scale features to maintain boundary details and detect completely changed targets. Finally, we use the simple yet efficient UNet structure to recover the feature maps and predict results. To demonstrate the effectiveness of the MSGFNet, the LEVIR-CD, WHU-CD, and SYSU-CD datasets were utilized, and the MSGFNet achieved F1 scores of 90.86%, 92.46%, and 80.39% on the three datasets, respectively. Furthermore, the low computational costs and small model size have validated the superior performance of the MSGFNet.

Keywords: change detection; remote sensing images; multi-scale progressive fusion; gated weight adaptive fusion

Citation: Wang, Y.; Wang, M.; Hao, Z.; Wang, Q.; Wang, Q.; Ye, Y. MSGFNet: Multi-Scale Gated Fusion Network for Remote Sensing Image Change Detection. *Remote Sens.* **2024**, *16*, 572. https://doi.org/10.3390/rs16030572

Academic Editors: Kamil Krasuski and Damian Wierzbicki

Received: 1 January 2024
Revised: 27 January 2024
Accepted: 29 January 2024
Published: 2 February 2024

1. Introduction

The advance of satellite imaging technology has facilitated the acquisition of remote sensing images (RSIs). Change detection (CD) is the process of identifying changes in the ground within the same geographical area utilizing RSIs taken at two different times [1]. Due to its wide application in urban sprawl detection [2], urban green ecosystems [3], damage assessment [4], etc., CD as a fundamental and important task has increasingly gained attention in the remote sensing field.

During the early stages of CD research, numerous methods have been proposed by researchers [5,6]. For example, image difference was one of the earliest CD methods for subtracting bi-temporal images according to the corresponding pixels [7]. To address spurious changes and counter positional errors, a robust change vector analysis method was proposed by Thonfeld et al. [8], combining intensity information with the advantages of change vector analysis (CVA). Researchers have made substantial progress through extensive research on these traditional methods [9–11]. However, these traditional CD methods face new challenges with the increased spatial resolution of remote sensing images. On one hand, traditional CD methods are designed for medium- and low-resolution RSIs, resulting in poor performance when dealing with rich information in high-resolution RSIs [12]. On the other hand, these methods rely on handcrafted features that are sensitive

to radiation differences and illumination changes [13,14]. Consequently, the application of traditional CD methods is limited in scope.

Recently, with the advent of the big data era, deep neural networks have demonstrated their strong feature extraction capabilities [15,16], with the end-to-end advantages of convolutional neural networks (CNNs) being particularly notable. CNNs have been widely employed in CD tasks and have spawned a number of promising CD methods [17,18]. For example, Zhang et al. [19] integrated a CycleMLP block into a Siamese network, proposing an MLP-based method for CD. However, it is important to note that this method incurs a substantial inference time. Fang et al. [20] introduced a CD method that combines the UNet++ architecture with a Siamese network. This method mitigates the loss of localization feature information by establishing a dense connection between the encoder and decoder.

Although the methods mentioned above have achieved performance results, they do not consider the characteristics of bi-temporal multi-scale features, thereby resulting in incomplete detection targets and limited accuracy of results. Inspired by the widely used multi-scale pyramid architecture for extracting multi-scale feature information in medical image segmentation [21], several methods have been proposed to address these problems by using multi-scale features [22–24]. For instance, Li et al. [23] proposed a multi-scale convolutional channel attention mechanism to generate detailed local features and integral global features. For capturing feature information on all scales, Xiang et al. [22] introduced a multi-receptive field position enhancement module incorporating convolutional layers with different kernel sizes. Despite the improvements achieved by the above methods through the incorporation of multi-scale features, they still exhibit certain shortcomings. On the one hand, these methods employ a simple concatenation strategy for fusing multi-scale features without considering the interaction between them. On the other hand, they extract multi-scale features after a simple feature fusion (i.e., feature difference) rather than employing bi-temporal multi-scale feature fusion. Consequently, the simple feature fusion often has restrictions that are not discriminative enough and result in unsmooth detection boundaries.

To address such problems, this study investigates the multi-scale fusion of bi-temporal features to detect complete change targets and improve the accuracy of results, and we further propose a multi-scale gated fusion network (MSGFNet). In particular, we opt for a lightweight model, namely EfficientNetB4 [25], as the encoder for constructing the Siamese architecture. This architecture is utilized to extract multi-layer features from bi-temporal images. Then, we propose a multi-scale gated fusion module (MSGFM) that has a multi-scale progressive fusion (MSPF) unit and a gated weight adaptive fusion (GWAF) unit. This module aims to obtain discriminative fusion features, improving the details of boundaries and effectively detecting the complete change targets. To gradually reconstruct the results, the decoder processes the fused multi-scale features in the end. The main contributions of this study may be summarized as follows:

1. We propose a novel end-to-end CD network, namely the multi-scale gated fusion network (MSGFNet). The MSGFNet is designed with a weight-sharing Siamese architecture tailored to be compatible with the CD task;
2. To improve the details of boundaries and detect the complete change targets, we propose an MSGFM comprising an MSPF unit and a GWAF unit. The MSGFM adaptively fuses bi-temporal multi-scale features based on gate mechanisms to obtain discriminative fusion features;
3. To confirm the efficacy of the MSGFNet, we employed the LEVIR-CD, WHU-CD, and SYSU-CD datasets for our comparison experiments. The results demonstrate that the MSGFNet outperforms several state-of-the-art (SOTA) methods. Additionally, the MSGFM was validated through ablation studies.

The following section outlines the organization of the remainder of the paper. A brief review of the latest relevant works is given in Section 2. Section 3 details the overall framework of the MSGFNet. Section 4 sequentially offers information on experimental datasets,

evaluation metrics, comparison methods, experimental details, results, and ablation studies. Section 5 concludes the paper.

2. Related Work

In this section, a brief review of the latest methods based on deep learning is given. The current deep-learning-based CD methods can be categorized into three groups based on network structure: CNN-based, transformer-based, and hybrid-based methods.

2.1. CNN-Based Methods

From the perspective of the fusion strategy, CNN-based methods can be further categorized into single-stream and two-stream methods [26]. In detail, single-stream methods take inspiration from semantic segmentation tasks. Researchers have proposed some approaches to image-level fusion strategies that match the semantic segmentation networks. For instance, Sun et al. [27] introduced conventional long short-term memory into Unet for CD. Peng et al. [28] employed bi-temporal images that had been concatenated into a UNet++ network. They further proposed a fusion strategy on multiple side outputs to improve the accuracy of results. Nevertheless, the independent feature characteristics of each bi-temporal image cannot be directly captured by single-stream CD methods based on semantic segmentation networks.

In contrast to single-stream, two-stream methods leverage the Siamese architecture, which consists of two streams that share weights to generate features of bi-temporal images. Most existing CD methods [20,29–31] adopt the Siamese architecture because it is appropriate for handling the input of RSIs. For instance, Dai et al. [29] introduced a building CD method that comprises a multi-scale joint supervision module and an improved consistency regularization module. Ye et al. [30] employed Siamese networks to propose a feature decomposition optimization reorganization network for CD. The edge and main body features were modeled using a feature decomposition strategy. Li et al. [32] proposed a lightweight CD method composed of three modules: a neighbor aggregation module (NAM), a progressive change identifying module (PCIM), and a supervised attention module (SAM), to improve the accuracy of results. Zhou et al. [33] introduced a context aggregation method utilizing a Siamese network. The multi-level features were fed into a context extraction module in this method, enabling the acquisition of long-range spatial-channel context features.

2.2. Transformer-Based Methods

Transformer-based methods, originally developed for natural language processing, are now being applied to encode bi-temporal images for CD. For example, Bandara et al. [34] introduced a CD method that combines a transformer with a Siamese architecture. This method introduced a transformer feature encoder to extract coarse and fine features with high and low resolution, respectively. Song et al. [35] introduced a progressive sampling transformer network (PSTNet) by using the excellent modeling ability of the transformer. In this method, the optimized tokens are iteratively mapped back to the original features to establish enhanced spatial connections in the spatial domain. Fang et al. [36] introduced a CD method, Changer, which uses a Siamese hierarchical transformer to extract multilayered features and then designs a flow-based dual-alignment fusion module to fuse the two branches' features. Zhang et al. [37] introduced a CD method that used a pure Swin transformer utilizing a Siamese network to extract long-term global features. However, transformer-based methods face limitations in terms of computational complexity and larger parameter sizes [38]. In addition, transformer-based methods often result in irregular boundaries in the results due to their disregard for the subtle details of shallow features.

2.3. Hybrid-Based Methods

Hybrid-based methods combine CNN and transformer architectures, which aim to improve feature extraction abilities [39]. For example, to couple the global and local fea-

tures, Feng et al. [40] integrated a transformer and a CNN to design a CD method that was composed of an inter-scale feature fusion module and an intra-scale cross-interaction module, which were designed for obtaining discrimination feature maps and constructing spatial–temporal contextual information, respectively. To address the issues of blurred edges and neglect caused by sampling that is either too shallow or too deep, Song et al. [41] introduced a simple convolutional network and a progressive sampling CNN to generate fine and coarse features, respectively. Subsequently, a mixed-attention module was introduced to merge coarse and fine features. Finally, the results were generated by feeding the fused features into a transformer decoder. Chu et al. [42] proposed a dual-branch feature-guided aggregation network for CD. This method employs a dual-branch structure composed of a CNN and s transformer to extract both semantic and spatial features at various scales. However, in this method, the feature extractor is not only complicated but the network also has a large number of parameters. Tang et al. [43] introduced a W-shaped dual Siamese network (WNet) for CD. In this method, a deformable convolution was introduced into the CNN branch and transformer to mitigate the limited receptive fields and regular patch generation, respectively. Similarly, this method also possesses a significant number of parameters. Moreover, hybrid-based CD methods further require the design of a complicated fusion module to fuse the CNN features and token features, which are extracted from the CNN network and transformer network, respectively.

3. Materials and Methods

3.1. Framework

As depicted in Figure 1, the MSGFNet follows a standard U-shaped [44] network that employs a Siamese architecture. In particular, the MSGFNet comprises a Siamese feature encoder, an MSGFM, and a decoder for result prediction. First, to preserve the independence of features in bi-temporal images [45], each bi-temporal image is separately fed into the shared-weight Siamese EfficientNetB4 to generate the multi-level features. Subsequently, to effectively fuse multi-scale features aimed at improving the details of changed boundaries, the MSGFM is designed to adaptively fuse the corresponding bi-temporal features at the same feature level. The fused features are decoded following the same skip connection method as in the classic UNet architecture [44], followed by a sigmoid classifier to generate the results.

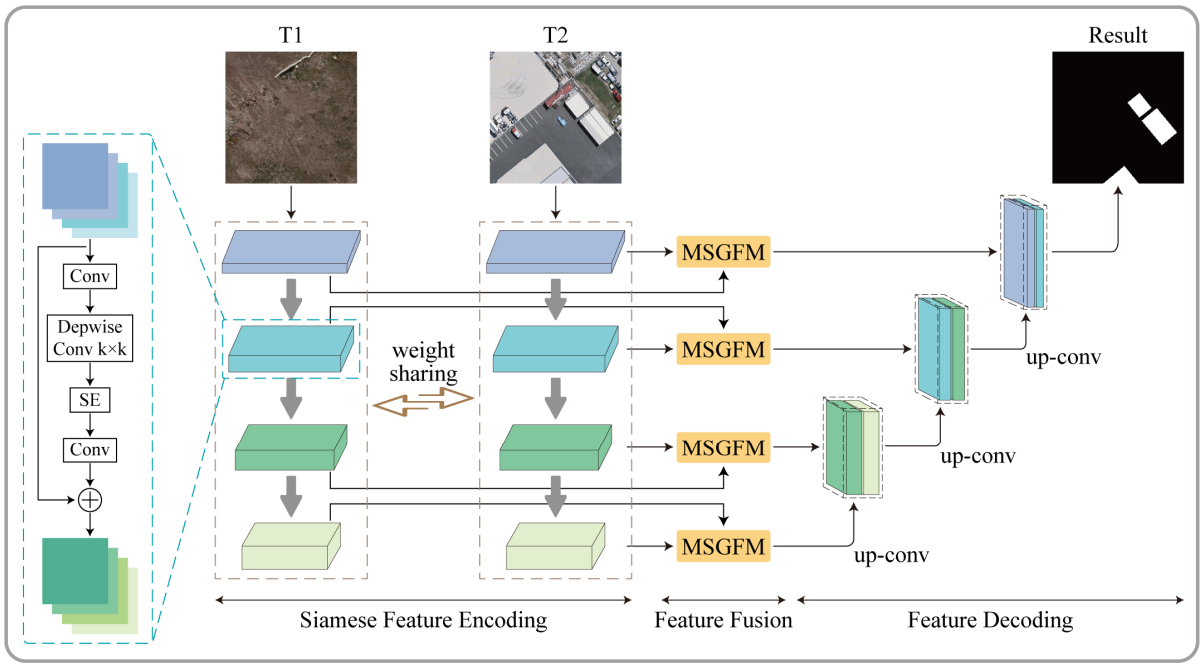

Figure 1. General overview of the MSGFNet.

3.2. Siamese Feature Encoder

Considering feature extraction abilities, network parameters, and computational memory, we chose EfficientNetB4 [25] as the backbone encoder for the Siamese architecture. More specifically, we made use of the first four convolutional stages of EfficientNetB4 that have been pre-trained on ImageNet. In particular, the first stage is a common 3×3 convolutional layer. The second, third, and fourth stages are each composed of identical MBConv blocks, with 2, 4, and 4 MBConv blocks, respectively.

The structure of the MBConv is depicted in Figure 2. In particular, the MBConv is composed of two 1×1 convolutional layers, a $k \times k$ depthwise convolutional layer, and a squeeze–excitation module. Within the MBConv block, the input features' channel dimension is increased by using the first convolutional layer. The kernel size k of the depthwise layer in the fourth stage is 5, whereas in other stages, it is 3. Squeeze–excitation is a specific attention mechanism that is able to suppress background feature information and enhance significant information. The purpose of the final convolutional layer is to reduce the channel dimension of the features to align them with the input features, allowing for the utilization of a residual connection mechanism. More details can be found in the literature [25].

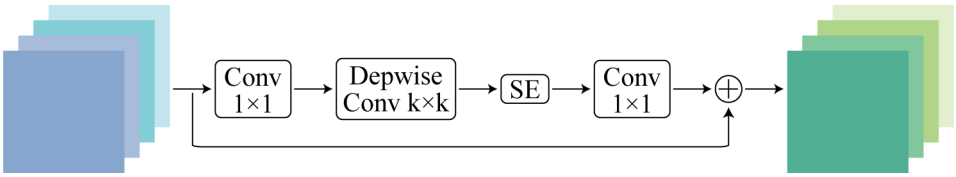

Figure 2. The structure of the MBConv block.

Given the bi-temporal images represented as $I^1, I^2 \in R^{C \times H \times W}$, where H, W, and C denote the height, width, and image band numbers, respectively, the bi-temporal images are then separately input into each branch corresponding to the first four stages of the Siamese EfficientNetB4 to generate multi-level features. As a result, the multi-level features are represented as $f_i^1, f_i^2, i \in \{1, 2, 3, 4\}$, respectively, where i represents the i-th stage. The feature depths of the four stages are 48, 24, 32, and 56, respectively. The spatial scales of the extracted multi-level features in the successive stages are $\left\{ \frac{H}{2} \times \frac{W}{2}, \frac{H}{2} \times \frac{W}{2}, \frac{H}{4} \times \frac{W}{4}, \frac{H}{8} \times \frac{W}{8} \right\}$.

3.3. Multi-Scale Gated Fusion Module

Generally, the changed objects in bi-temporal images often have significant size variations [24], which leads to incomplete detection targets and unsmooth boundaries in the results. Consequently, it is imperative to explore multi-scale feature fusion strategies to smooth the boundaries and improve the accuracy of results. Hence, an MSGFM that is capable of adaptively fusing multi-scale features is proposed. More specifically, the MSGFM comprises an MSPF unit and a GWAF unit.

3.3.1. Multi-Scale Progressive Fusion Unit

Previous studies [46,47] have demonstrated that the local receiving field is insufficient for accurately detecting ground objects of various shapes and sizes. To better capture ground objects of different sizes, we propose the use of an MSPF unit (Figure 3). Specifically, there are four parallel atrous convolutions and a progressive connection strategy used by the MSPF unit to progressively fuse the multi-scale features.

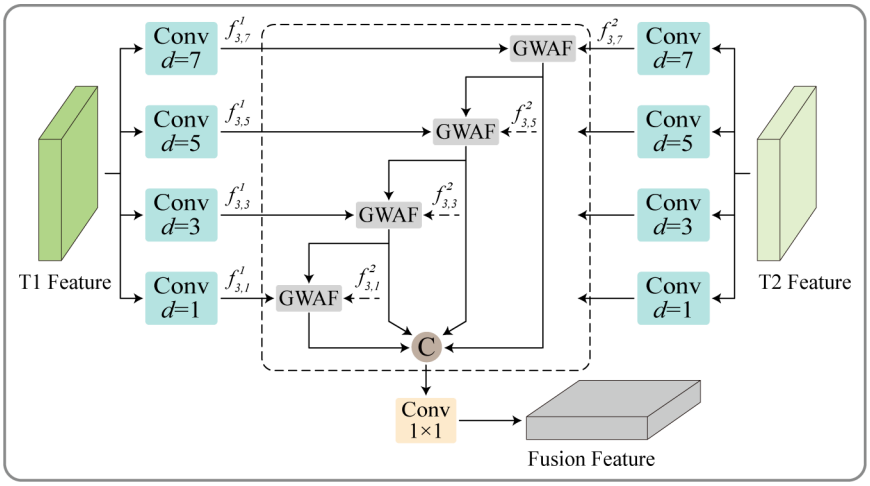

Figure 3. The structure of the proposed MSPF unit.

Consider a pair of bi-temporal features of any stage generated from a Siamese feature encoder, denoted as f^1 and f^2. First, to effectively capture multi-scale feature information about ground objects, we utilize four parallel atrous convolutions with the same kernel size but different atrous rates to generate features at different pyramid scales. In particular, the kernel size for all four convolutions is set to 3×3, and the atrous rates of the four convolutions are 7, 5, 3, and 1, respectively. In addition, the output channels of the four convolutions are set to one-fourth of the channel of the input features. For instance, the bi-temporal features f^1 and f^2 are inputted into the four parallel atrous convolutions, which can be denoted as follows:

$$\begin{cases} f_{3,i}^1 = Conv_{3,i}\left(f^1\right) \\ f_{3,i}^2 = Conv_{3,i}\left(f^2\right) \end{cases} i \in \{7,5,3,1\} \tag{1}$$

where $Conv_{3,i}$ is the convolution function with different atrous rates, the subscript 3 denotes the kernel size of the convolution, and the subscript i represents the atrous rate of each convolution. $f_{3,i}^1$ and $f_{3,i}^2$ are the pyramid features, respectively.

As described above, the proposed MSPF is a progressive process proposed to fuse bi-temporal multi-scale features. Specifically, the features $f_{3,7}^1$ and $f_{3,7}^2$ are fed into the GWAF to achieve weighted adaptive feature fusion. In addition, to mitigate the loss of fused feature information, each set of fused features using the upper GWAF is inputted into the next GWAF based on a progressive connection strategy, as depicted in Figure 3. The specifics of the GWAF will be explained in the next section. The above process can be formulated as follows:

$$\begin{cases} f_7 = GWAF\left(f_{3,7}^1, f_{3,7}^2\right) \\ f_i = GWAF\left(f_{3,i}^1, f_{3,i}^2, f_{i+2}\right) \end{cases} i \in \{5,3,1\} \tag{2}$$

where $GWAF$ denotes the weighted adaptive fusion operation and f_i is the fused feature for each scale. It is essential to note that the first GWAF fusion unit does not have a progressive connection input. Subsequently, the four fused features are concatenated along the channel dimension, followed by a 1×1 convolutional layer employed to produce the discriminative fusion features. The process is formulated as follows:

$$F_i = Conv_1([f_1; f_3; f_5; f_7]) \; i \in \{1,2,3,4\} \tag{3}$$

where F_i represents the fused multi-level features and $Conv_1$ denotes a 1×1 convolutional layer.

3.3.2. Gated Weight Adaptive Fusion Unit

Previous studies [39,48,49] have generally fused the bi-temporal features using simple summation or concatenation. Nevertheless, it is difficult for these direct fusion strategies to effectively highlight the changed feature information and suppress the unchanged feature information. Taking inspiration from the gate mechanism [50], which can learn to highlight the contributions of changed regions, we propose a GWAF unit for bi-temporal multi-scale feature weighted adaptive fusion. Figure 4 depicts the details of the GWAF unit.

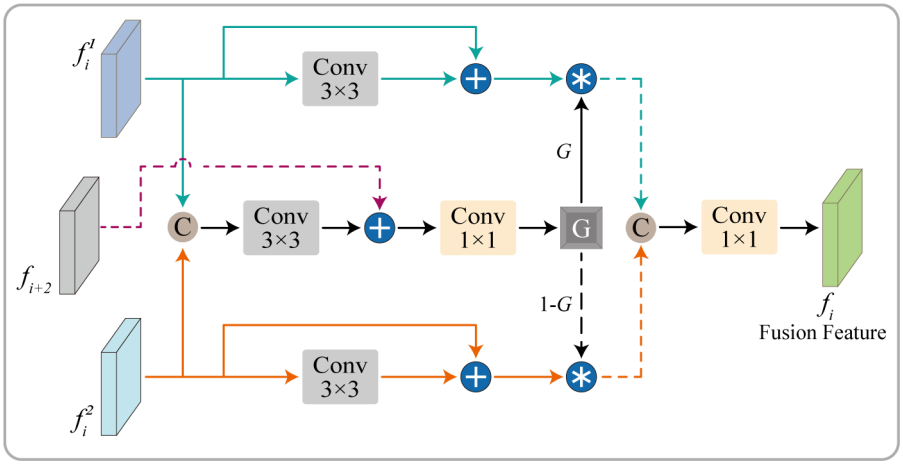

Figure 4. The structure of the GWAF unit.

Given the same scale, bi-temporal features are represented as f_i^1 and f_i^2, $i \in \{7,5,3,1\}$. In particular, the GWAF unit is roughly composed of three branches, the individual inputs (i.e., f_i^1 and f_i^2) generated from the multi-scale atrous convolutional layer, and one fused feature (f_{i+2}) obtained from the upper-scale GWAF unit. It is essential to note that the top-scale GWAF unit does not contain the additional fused feature. For convenient illustration, we simplified the subscripts of symbols.

To obtain the gated weight map G_i between bi-temporal features, the bi-temporal features f_i^1 and f_i^2 are first concatenated along the channel dimension, and then a 3×3 convolutional layer is used to fuse the bi-temporal features. Subsequently, the fused feature f_{i+2} obtained from the upper-scale GWAF unit is added to the current scale with a residual connection strategy. After that, a sigmoid function is applied after a 1×1 convolutional layer to further fuse the multi-scale feature information to obtain the gated weight map G_i. The process can be formulated as follows:

$$f_i^{cat} = Conv_3\left(\left[f_i^1; f_i^2\right]\right) \tag{4}$$

$$G_i = Conv_1\left(f_{i+2} + f_i^{cat}\right) \tag{5}$$

where f_i^1 and f_i^2 are the bi-temporal features, $Conv_3$ is a 3×3 convolutional layer, f_i^{cat} denotes the concatenation features, and f_{i+2} is the fused feature generated from the upper-scale GWAF unit. $Conv_1$ is the function of 1×1 convolution followed by a sigmoid layer. G_i denotes the gated weight map.

To use the gated weight map G_i to refine the changed feature information, the feature f_i^1 is inputted into a 3×3 convolutional layer to extract more abstract semantic feature information. Then, the residual connection strategy is employed to add the features before and after convolution. Subsequently, the gated weight map G_i is element-wise multiplied with the newly added features to generate the discriminate fused features. In addition, the

inverse gated weight map $1 - G_i$ is used for f_i^2 to generate the enhanced feature f_G^2. The process can be formulated as follows:

$$\begin{cases} f_G^1 = G_i * \left(f_i^1 + Conv_3\left(f_i^1\right)\right) \\ f_G^2 = (1 - G_i) * \left(f_i^2 + Conv_3\left(f_i^2\right)\right) \end{cases} \tag{6}$$

where f_G^1 and f_G^2 represent the adaptively fused bi-temporal features corresponding to f_i^1 and f_i^2, respectively. After that, the enhanced features are concatenated along the channel dimension. Finally, a 1×1 convolutional layer is utilized to obtain the fused features of the i-th GWAF unit. The above process can be formulated as follows:

$$f_i = Conv_1\left(\left[f_G^1; f_G^2\right]\right) \tag{7}$$

where f_i represents the features generated from the i-th GWAF unit. By combining the GWAF unit with the MSPF unit, this approach is capable of efficiently fusing the bi-temporal features to highlight the changed feature while suppressing the unchanged feature in bi-temporal images.

3.4. Decoder

A decoder is employed to reconstruct the multi-level fused features to produce the results [23]. UNet is a widely used semantic segmentation network that uses skip connections to transmit detailed feature information from the encoder to the decoder [44]. As a result, many researchers have incorporated UNet into CD tasks and proposed a series of CD methods [51–53]. Following this, we use UNet, which has a simple yet effective architecture, to generate the change maps.

Generally, the fourth-level fused bi-temporal features (i.e., F_4) are up-sampled to the spatial size of the third-level fused features (i.e., F_3). After that, the features up-sampled from F_4 and F_3 are concatenated in the feature direction. Subsequently, a convolutional block is utilized to project the concatenated features to obtain the corresponding features with the same channel numbers as the F_3. The convolutional block comprises a 3×3 convolutional layer, a BN layer, and a ReLU layer. The above process can be formulated as follows:

$$\overline{F}_3 = ReLU(BN(Conv_3([F_3; Up_2(F_4)]))) \tag{8}$$

where Up_2 is the up-sampled operation and $\overline{F}_3$ represents the generated features that have the same channel numbers as F_3 The above steps are repeated until we obtain the features $\overline{F}_1$ Finally, the last 1×1 convolutional layer is employed to map the features $\overline{F}_1$ to the predicted maps, which have two channels (i.e., representing the changed and unchanged classes).

3.5. Details of Loss Function

In classification tasks, the cross-entropy (CE) function is frequently employed, and the CD can be regarded as a unique two-label classification. Consequently, the loss function is the CE function, which is employed as the loss function and expressed as follows:

$$L_{ce} = -\frac{1}{H \times W} \sum_{i=1}^{H \times W} \left[t_i \log(p_i) + (1 - t_i) \log(p_i)\right] \tag{9}$$

$$p_i = \begin{cases} \overline{p} & \text{if } t_i = 1 \\ 1 - \overline{p} & \text{otherwise} \end{cases} \tag{10}$$

where H and W are the image height and weight, respectively, $\overline{p}$ is the probability of the prediction results, and t_i represents the corresponding truth map.

4. Results

The three public building datasets that were used in the experiments are first described. Next, the evaluation metrics, comparison methods, and experimental details are introduced in turn. Finally, the results of the experiment are carefully investigated.

4.1. Datasets

4.1.1. WHU-CD

The WHU-CD dataset [54] is a dataset for detecting building changes. This dataset contains a pair of aerial images with $32,507 \times 15,354$ pixels that were obtained in 2012 and 2016, respectively, and has a spatial resolution of 0.2 m. The bi-temporal images in this study were cropped into 256×256 non-overlapping sub-images (Figure 5a), which were then divided into training, validation, and testing at an 8:1:1 ratio.

4.1.2. LEVIR-CD

The LEVIR-CD dataset [55] is a large-scale CD dataset that was collected using Google Earth from 2002 to 2018. This dataset includes 637 pairs of the size 1024×1024, with a spatial resolution of 0.5 m. Limited to the graphics processing unit (GPU) memory and following the division setting of the official study [55], the bi-temporal images were cropped into 256×256 non-overlapping sub-images (Figure 5b), and 7120/1024/2048 pairs were obtained for training, validation, and testing.

4.1.3. SYSU-CD

The SYSU-CD dataset [56] comprises a total of 20,000 pairs of aerial images with a resolution of 0.5 m and a spatial size of 256×256. This dataset encompasses various change types occurring in complex scenarios, such as building dilation, vegetation change, and sea construction. Following the official settings [56], the pairs in this dataset were divided into training, validation, and testing, with 12,000, 4000, and 4000 pairs, respectively. Some examples are illustrated in Figure 5c.

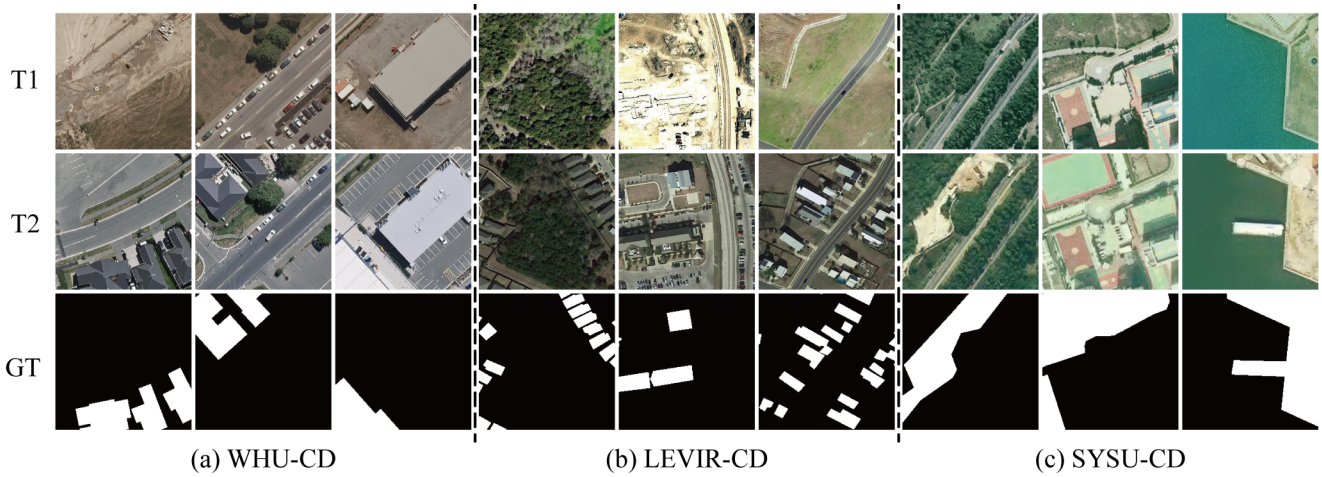

Figure 5. Example display of the three datasets, where T1 and T2 present bi-temporal images and GT represents the ground truth.

4.2. Evaluation Metrics

The four common evaluation measures that we used to thoroughly assess the MS-GFNet were precision, recall, F1, and intersection over union (IoU). In these evaluation indicators, precision and recall denote detection error and omission error, respectively. F1 is a more comprehensive metric that could be computed by taking the harmonic mean of

recall and precision [17]. Therefore, this paper selects F1 and IoU as the main evaluation measures. These measures described above are defined as follows:

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

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

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

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

where TP represents the pixel numbers of true positives, TN represents the pixel numbers of true negatives, FN represents the pixel numbers of false negatives, and FP denotes the pixel numbers of false positives.

4.3. Comparison Methods

We conducted a comparative analysis using eight SOTA CD methods to assess the performance of the MSGFNet. A brief description of these methods is provided below:

1. FC-EF [57]: FC-EF stands as a milestone method, utilizing a classic U-Net architecture. In this method, the bi-temporal images are concatenated along the feature direction before being input into the network.
2. FC-Siam-Diff [57]: FC-Siam-diff is a CD method with a Siamese CNN architecture. This network first extracts multi-level features from bi-temporal images and then uses the feature difference as the feature fusion module to generate change information.
3. STANet [55]: STANet is a metric-based method. This method suggests using a spatiotemporal attention module based on self-attention mechanisms to model the spatial and temporal relationships to obtain significant information about changed features.
4. DSIFNet [58]: DSIFNet is a deeply supervised image fusion method. This method proposes an attention module to integrate multilevel feature information and employs the deep supervision strategy to optimize the network and improve its performance.
5. SNUNet [20]: SNUNet is a combination of the NestedUNet and Siamese networks. This method alleviates the localization information loss by using a dense connection between the encoder and decoder. Furthermore, an ensemble channel attention module is built to refine the change features at different semantic levels.
6. BITNet [59]: BITNet is a combination of a transformer and a CNN. This network first extracts semantic features by using the CNN, and then uses the transformer to model the global feature into a set of tokens, strengthening the contextual information of the changed features.
7. ChangeFormer [34]: ChangeFormer is a purely transformer-based change detection method. This method uses a Siamese transformer to build the bi-temporal image features and then uses the multi-layer perceptual to decode the difference features.
8. LightCDNet [60]: LightCDNet employs a lightweight MobileNetV2 to extract multilevel features and introduces a multi-temporal feature fusion module to fuse the corresponding level features. Finally, deconvolutional layers are utilized to recover the change map.

To achieve a fair comparison, all the different comparison methods were evaluated under the same experimental setting. If the comparison methods and the proposed MSGFNet used the same dataset, we utilized the pre-trained weight models provided by the respective comparison papers. Otherwise, we employed the provided code and default parameters of the comparison methods.

4.4. Experimental Details

In this study, NVIDIA GeForce RTX 3080Ti graphics cards with 12 GB of RAM were used for training and all experiments were carried out using the PyTorch framework. During the training process, the AdamW optimizer was employed with a weight decay equal to 1×10^{-4}, and an initial learning rate of 1×10^{-3}. In addition, all experiments utilized a batch size of 8 and each dataset underwent training for 100 epochs.

4.5. Results

We analyzed the MSGFNet using the six SOTA methods on the two datasets in this part. We categorized the six SOTA methods into three classes: CNN-based, transformer-based, and hybrid-based. To enhance readability in the visualization comparisons, we depict the false positives in red, the false negatives in green, the true negatives in black, and the true positives in white.

4.5.1. Experimental Analysis on the WHU-CD Dataset

The experimental results on the WHU-CD dataset are displayed in Table 1. Notably, it can be observed that the proposed MSGFNet shows outstanding performance, with an F1 of 92.46% and an IoU of 85.98%. Furthermore, the hybrid-based BITNet and transformer-based ChangeFormer secure the second and third positions, respectively, with F1 scores of 91.25% and 89.82%. The proposed MSGFNet exhibits a superior F1 compared to BITNet and ChangeFormer, surpassing them by 1.21% and 2.64%, respectively. In addition, the BITNet and ChangeFormer outperform other CNN-based methods except for LightCDNet. However, despite the utilization of a CNN-based architecture, the MSGFNet achieved optimal results compared to the SOTA methods. This can be attributed to the effectiveness of the proposed MSGFM in capturing discriminatively changed feature information between the bi-temporal images.

Table 1. Quantitative evaluation of the MSGFNet and the SOTA methods on the WHU-CD dataset. The best scores are marked in **Bold**.

Methods		Pre	Recall	F1	IoU
CNN-based	FC-EF	79.33	74.58	76.88	62.45
	FC-Siam-Diff	67.55	63.21	65.31	48.75
	STANet	86.11	88.14	87.11	77.17
	DSIFNet	85.89	91.31	88.52	79.40
	SNUNet	82.63	90.33	86.31	75.92
	LightCDNet	92.00	91.00	91.50	84.30
Transformer-based	ChangeFormer	89.36	90.28	89.82	81.60
Hybrid-based	BITNet	**92.71**	89.83	91.25	84.30
CNN-based	MSGFNet	91.88	**93.06**	**92.46**	**85.98**

An intuitive visual comparison of all the methods is shown in Figure 6. It can be observed that both the STANet and FC-Siam-Diff not only exhibit a significant number of false negatives but also have rough boundaries in changed regions. Additionally, the boundary detection results reported by both the SNUNet and DSIFN are unsatisfactory. Compared to the second-ranked LightCDNet, the MSGFNet not only has more accurate boundary details but also has few false positives and false negatives. In summary, the MSGFNet achieves the best visualization performance on the WHU-CD dataset.

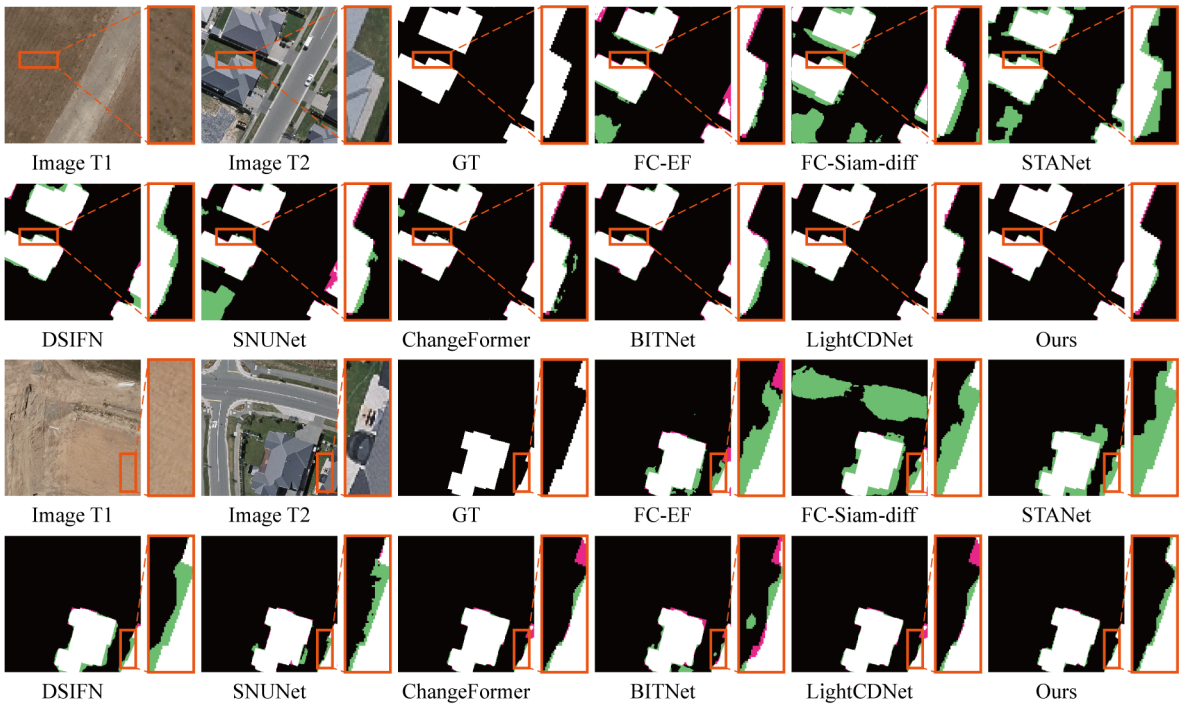

Figure 6. Visual comparisons between the MSGFNet and the SOTA methods on the WHU-CD dataset.

4.5.2. Experimental Analysis on the LEVIR-CD Dataset

The quantitative results of all methods on the LEVIR-CD dataset are displayed in Table 2. From the table, it is evident that the FC-Siam-Diff obtained the poorest performance. This may be attributed to the utilization of a simple Siamese UNet in the FC-Siam-Diff model; consequently, which leads to poor feature extraction and fusion ability. Correspondingly, other CNN-based methods, such as STANet, DSIFNet, and SNUNet, introduce various attention mechanisms that enhance the discriminative features of bi-temporal images. Consequently, these methods have shown varying degrees of improvement in the accuracy of results. ChangeFormer obtained the second-highest level of performance, achieving F1 and IoU of 90.40% and 82.48%, respectively. The MSGFNet has demonstrated improvements of roughly 0.46% in F1 and 0.77% in IoU when compared to ChangeFormer. In addition, the proposed method also achieved the highest precision, with a score of 92.12%. In conclusion, the quantitative analysis presented above validates the effectiveness of the MSGFNet.

Table 2. Quantitative evaluation of the MSGFNet and the SOTA methods on the LEVIR-CD dataset. The best scores are marked in **Bold**.

Methods		Pre	Recall	F1	IoU
CNN-based	FC-EF	85.87	82.22	83.35	72.43
	FC-Siam-Diff	88.59	80.72	85.37	74.48
	STANet	83.81	**91.00**	87.30	77.40
	DSIFNet	87.30	88.57	88.42	78.09
	SNUNet	90.55	89.28	89.91	81.67
	LightCDNet	91.30	88.00	89.60	81.20
Transformer-based	ChangeFormer	92.05	88.80	90.40	82.48
Hybrid-based	BITNet	89.24	89.37	89.31	80.68
CNN-based	MSGFNet	**92.12**	89.63	**90.86**	**83.25**

Figure 7 shows an intuitive visual comparison of all the methods on the LEVIR-CD dataset. For the first densely built case, the changed buildings in the results of the FC-Siam-

Diff, STANet, DSIFNet, and SNUNet are clustered together to some extent. The results of the proposed MSGFNet show better detail and boundaries for the small ground objects. For the case featuring buildings of different scales in Figure 7, there are illumination changes and building shadows present between the bi-temporal images. The results of the proposed MSGFNet show that it has fewer false positives and false negatives than several SOTA methods while also preserving the integrity of small ground targets.

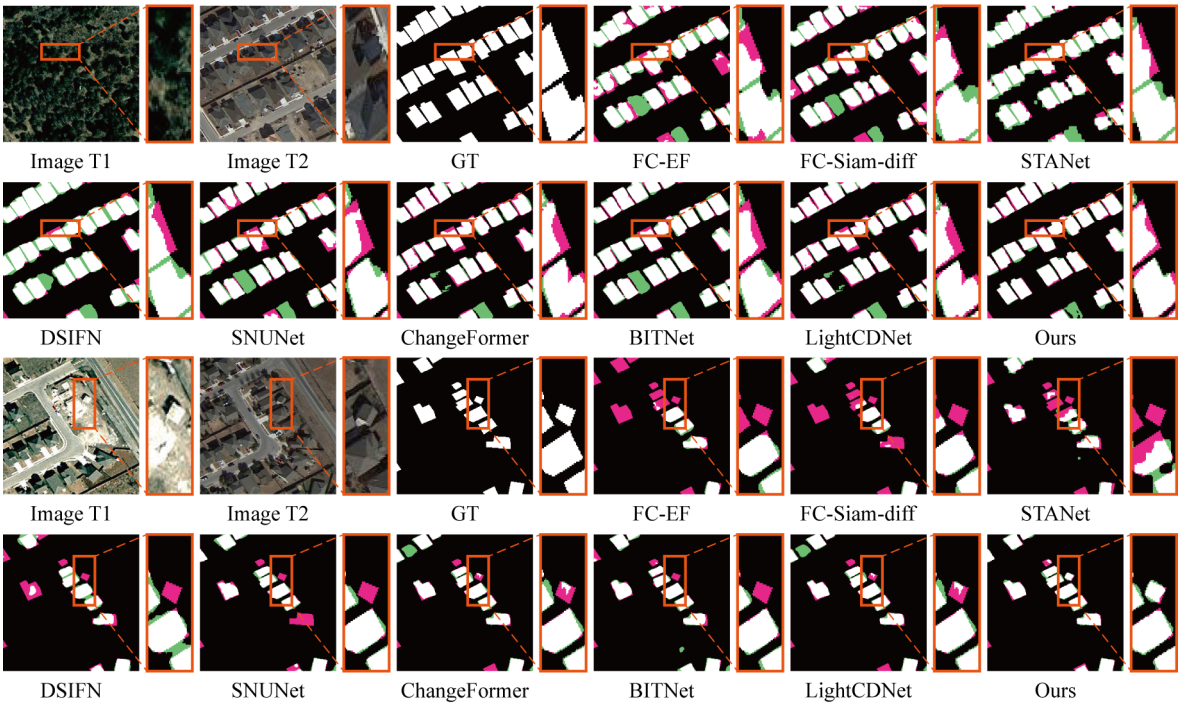

Figure 7. Visual comparisons between the MSGFNet and the SOTA methods on the LEVIR-CD dataset.

4.5.3. Experimental Analysis on the SYSU-CD Dataset

The experimental results on the SYSU-CD dataset are displayed in Table 3. Notably, the FC-Siam-Diff exhibits the least favorable performance, with an F1 value of 70.17% and an IoU of 55.11%. SNUNet slightly outperforms FC-Siam-Diff, which may be attributed to SNUNet's employment of the dense connection strategy that can alleviate the loss of feature information [20]. Among these comparative methods, STANet, DSIFNet, and ChangeFormer exhibit comparable performances. Specifically, the above three methods obtained F1 scores of 77.75%, 77.46%, and 77.83%, respectively. Correspondingly, LightCD-Net and BITNet were the second- and third-ranked methods, with F1 values of 78.52% and 78.72%, respectively. It is evident that the proposed MSGFNet outperforms the comparative methods in all evaluation metrics, except recall. Specifically, the proposed MSGFNet outperforms the second-ranked LightCDNet method by over 1.64% in F1. Despite the fact that STANet obtains the highest recall value, its F1 and IoU values are 2.64% and 3.63% lower than those of the proposed MSGFNet, respectively. In conclusion, the quantitative analysis presented above validates the effectiveness of the MSGFNet.

Table 3. Quantitative evaluation of the MSGFNet and the SOTA methods on the SYSU-CD dataset. The best scores are marked in **Bold**.

Methods		Pre	Recall	F1	IoU
CNN-based	FC-EF	80.16	70.69	75.13	60.17
	FC-Siam-Diff	78.34	66.13	70.17	55.11
	STANet	73.33	**82.73**	77.75	63.59
	DSIFNet	79.32	73.85	77.46	62.94
	SNUNet	82.16	71.33	76.36	61.76
	LightCDNet	83.01	74.90	78.75	64.98
Transformer-based	ChangeFormer	77.16	78.51	77.83	63.71
Hybrid-based	BITNet	80.40	77.09	78.72	64.90
CNN-based	MSGFNet	**83.34**	77.65	**80.39**	**67.22**

An intuitive visual comparison of all the methods is shown in Figure 8. Different from the WHU-CD and LEVIR-CD datasets, which only contain building changes, the SYSU-CD dataset is more challenging because it encompasses various change types occurring in complex scenarios [56]. For the building changes in the first case, the results of FC-EF and FC-Siam-Diff contain many missed detections (e.g., false negatives). However, the results of other comparative methods, such as DSIFNet and ChangeFormer, have many false detections (e.g., false positives). For the second case, which is a vegetation change sample, all the comparative methods have a large area of missed detection. For the two different change cases, compared to the comparative SOTA methods, only the proposed MSGFNet could detect the complete change ground objects and has the best visualization. Compared to the second-ranked LightCDNet method, the proposed MSGFNet not only has few false positives and false negatives but also maintains better boundary details. In summary, our method achieves optimal performance on the SYSU-CD dataset.

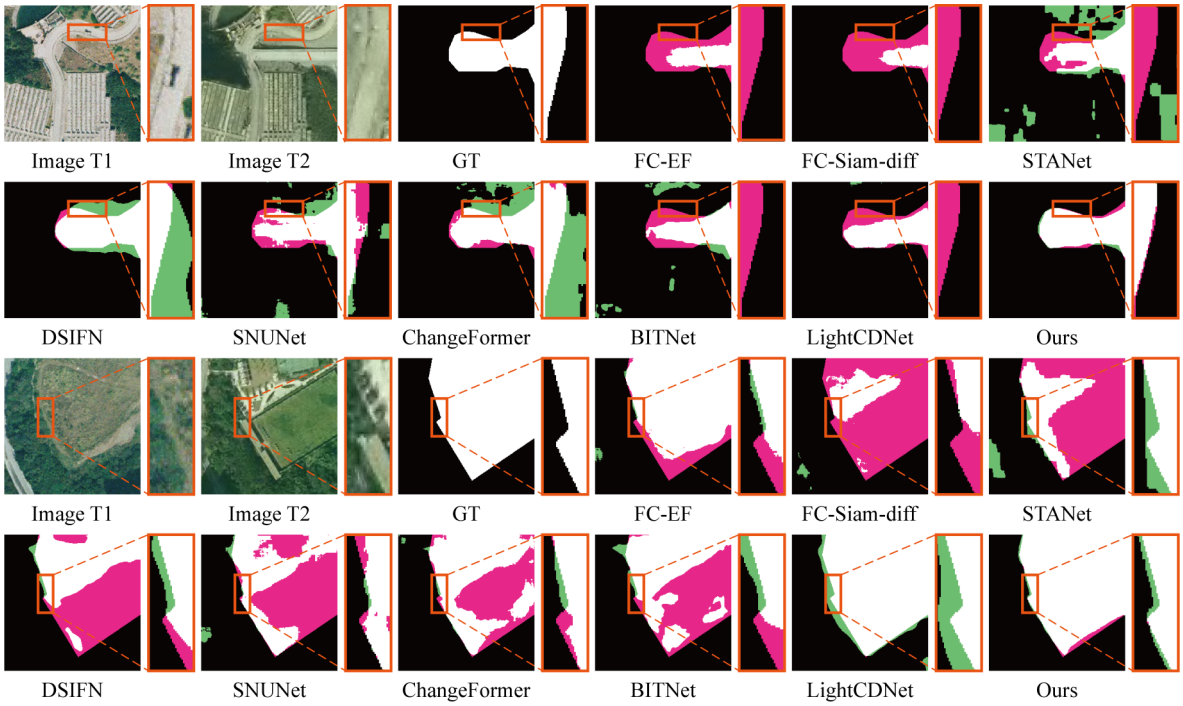

Figure 8. Visual comparisons between the MSGFNet and the SOTA methods on the SYSU-CD dataset.

4.5.4. Model Size and Computational Complexity

On the other hand, we conducted a comparative analysis of model size (number of parameters) and computational efficiency (number of floating-point operations) of all

the methods, as presented in Table 4. It is evident that the MSGFNet not only has the best performance in terms of F1 but also has the smallest model size in terms of network parameters. Specifically, the model parameters of the MSGFNet are just 0.58 M, which is lower than the FC-Siam-Diff and BITNet methods. Additionally, our method has the smallest FLOPs. The model size and computational complexity demonstrate that the MSGFNet more successfully obtains a compromise between performance and model size. For an intuitive visualization, the scatterplot between parameters and F1 of all methods is shown in Figure 9.

Table 4. Comparison of model size and computational complexity on the WHU-CD dataset.

Methods		Params/M	FLOPs/G	F1
CNN-based	FC-EF	1.35	3.58	76.88
	FC-Siam-Diff	1.35	4.73	65.31
	STANet	16.89	6.43	87.11
	DSIFNet	50.46	50.77	88.52
	SNUNet	12.03	54.83	86.31
	LightCDNet	10.75	21.54	91.50
Transformer-based	ChangeFormer	29.75	21.18	89.82
Hybrid-based	BITNet	3.04	8.75	91.25
CNN-based	MSGFNet	0.58	3.99	92.46

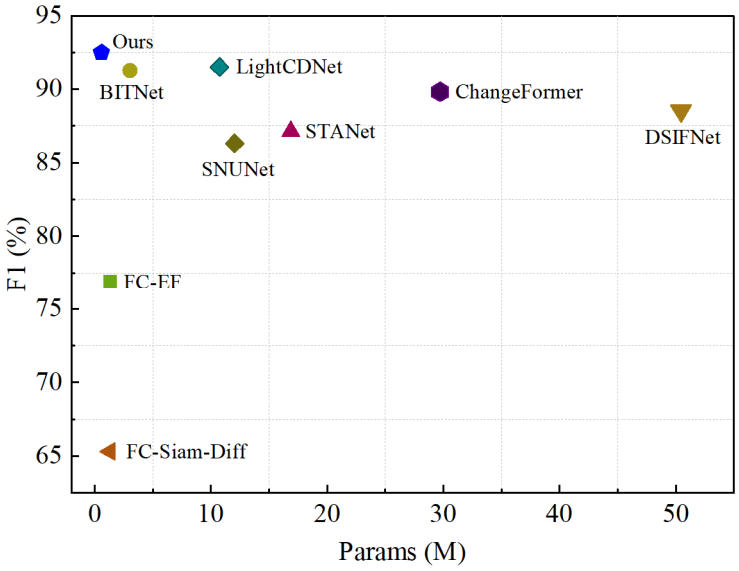

Figure 9. The scatterplot between parameters and F1 of all methods.

4.6. Ablation Studies

We conducted ablation studies on the LEVIR-CD dataset to demonstrate the effectiveness of the MSGF. Specifically, the proposed multi-scale gated fusion module consists of two units: a multi-scale progressive fusion unit and a gated weight fusion unit. Therefore, we performed individual corresponding ablation studies on both units. To begin with, "Base" refers to a Siamese encoder in the absence of any further modules, the multi-scale progressive fusion unit is denoted as "MSPF", and the gated weight fusion unit is represented as "GWAF". More specifically, we removed the MSPF unit to validate its effectiveness. In this scenario, we utilized only the GWAF to fuse the bi-temporal features. It is essential to point out that there is no additional input branch generated from the upper-scale GWAF unit. We replaced the GWAF unit with the general difference fusion mode to validate the effectiveness of the GWAF unit. In addition, we set up an additional control group. In this

control group, the network employs the same architecture as the FC-Siam-Diff to produce the change map.

Table 5 lists the quantitative results of the ablation studies. It can be observed that the mode "Base" without the MSPF and GWAF units generates the lowest performance. The results generated from each GWAF and MSPF unit are significantly better than the "Base" mode. Specifically, the utilization of the GWAF unit results in a 1.99% improvement in F1 and a 2.66% improvement in IoU. With the help of the MSPF unit, F1 and IoU are enhanced by 2.71% and 4.40%, respectively. Furthermore, the best results are produced when both the GWAF and MSPF units are combined at the "Base". In particular, there is an enhancement of 1.35% in F1 and 2.78% in IoU when the GWAF unit is added. On the other hand, when using the MSPF unit alone, F1 and IoU are both improved, by 0.63% and 1.04%, respectively. In general, these improvements indicate the effectiveness of the proposed GWAF and MSPF units.

Table 5. Quantitative evaluation results of ablation studies on the LEVIR-CD dataset. The best scores are marked in **Bold**.

Methods	Pre	Recall	F1	IoU
Base	90.14	85.05	87.52	77.81
Base + GWAF	90.66	88.39	89.51	80.47
Base + MSPF	91.69	88.83	90.23	82.21
Base + MSPF + GWAF	**92.12**	**89.63**	**90.86**	**83.25**

Some examples of the results are shown in Figure 10. It is evident that the results of the "Base" mode have many false positives and false negatives. When the GWAF and MSPF units are added, respectively, the results improve slightly. Furthermore, when the GWAF and MSPF are both added, we achieve optimal visualization results. Specifically, the results have fewer false negatives and false positives, and the boundary details are more precise. The visual results validate the effectiveness of the GWAF and MSPF units.

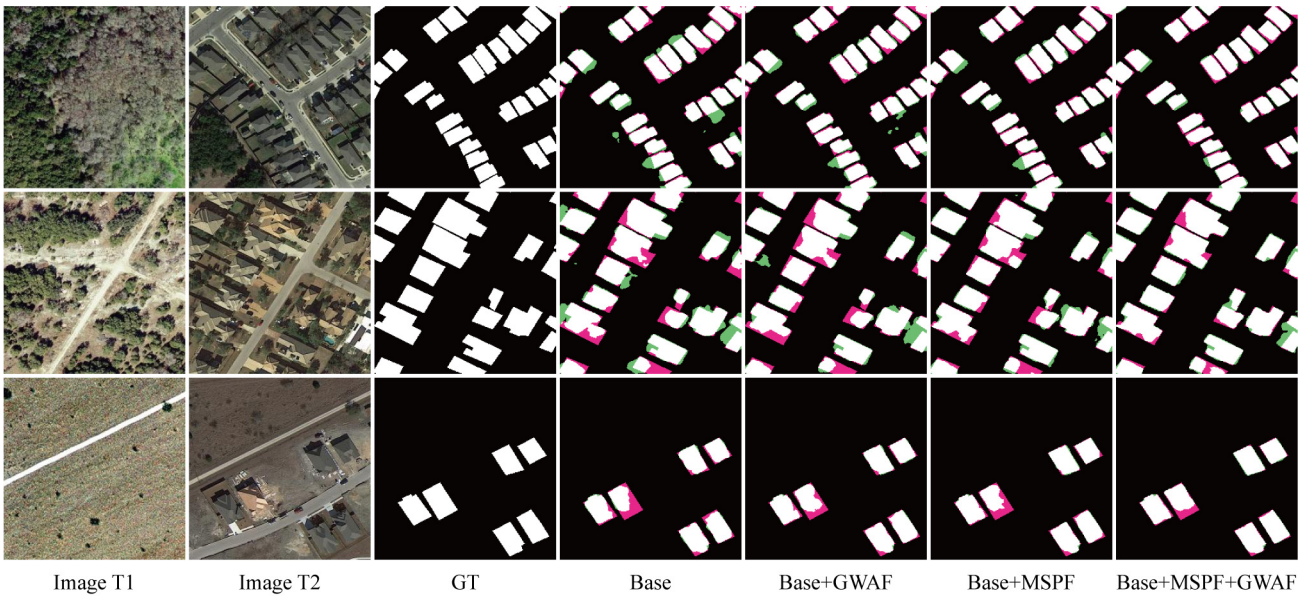

Figure 10. Visual comparisons of ablation experiments on the LEVIE-CD datasets.

5. Conclusions

This paper proposes a CD method, namely the MSGFNet. To capture useful feature information, the MSGFNet combines EfficientNetB4 with a Siamese structure to extract the multi-level features. An MSGFM that comprises an MSPF unit and a GWAF unit

is proposed to progressively and adaptively fuse bi-temporal multi-scale features. This module can obtain discriminative fusion features to smooth the details of changed object boundaries and improve the accuracy of results. Finally, the results obtained from three publicly available datasets show that the MSGFNet outperforms several SOTA methods in terms of both effectiveness and complexity. On the WHU-CD, LEVIR-CD, and SYSU-CD datasets, the MSGFNet achieved improvements of 1.21%, 0.46%, and 1.64% in F1 and 1.68%, 0.77%, and 2.24% in IoU, respectively, compared to the SOTA methods that produced better values. Additionally, it is evident that the Params and FLOPs for the proposed MSGFNet are 3.99 G and 0.58 M, respectively. Both values are lower than those of several SOTA methods. In summary, the proposed MSGFNet outperforms several SOTA CD methods.

Author Contributions: Conceptualization, M.W. and Y.Y.; methodology, Y.W., M.W. and Y.Y.; validation, Z.H., Q.W. (Qiang Wang) and Q.W. (Qianwen Wang); writing—original draft preparation, M.W. and Y.Y.; writing—review and editing, Y.W.; visualization, Z.H., Q.W. (Qiang Wang) and Q.W. (Qianwen Wang). All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (no. 42271446 and No. 41971281), and the Tianjin Key Laboratory of Rail Transit Navigation Positioning and Spatiotemporal Big Data Technology (no. TKL2023A12).

Data Availability Statement: The WHU-CD, LEVIR-CD, and SYSU-CD datasets are openly available at http://gpcv.whu.edu.cn/data/building_dataset.html (accessed on 15 October 2023), https://justchenhao.github.io/LEVIR/ (accessed on 15 October 2023), and https://github.com/liumency/SYSU-CD (accessed on 20 January 2024), respectively.

Acknowledgments: The authors are grateful to researchers for creating and providing publicly available datasets. Furthermore, the authors would like to express their gratitude to the anonymous reviewers for their valuable comments and suggestions.

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

References

1. Lu, D.; Mausel, P.; Brondízio, E.; Moran, E. Change Detection Techniques. *Int. J. Remote Sens.* **2004**, *25*, 2365–2401. [CrossRef]
2. Huang, X.; Zhang, L.; Zhu, T. Building Change Detection from Multitemporal High-Resolution Remotely Sensed Images Based on a Morphological Building Index. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 105–115. [CrossRef]
3. Shi, Q.; Liu, M.; Marinoni, A.; Liu, X. UGS-1m: Fine-Grained Urban Green Space Mapping of 31 Major Cities in China Based on the Deep Learning Framework. *Earth Syst. Sci. Data* **2023**, *15*, 555–577. [CrossRef]
4. Gao, F.; Dong, J.; Li, B.; Xu, Q.; Xie, C. Change Detection from Synthetic Aperture Radar Images Based on Neighborhood-Based Ratio and Extreme Learning Machine. *J. Appl. Remote Sens.* **2016**, *10*, 046019. [CrossRef]
5. Fang, H.; Du, P.; Wang, X.; Lin, C.; Tang, P. Unsupervised Change Detection Based on Weighted Change Vector Analysis and Improved Markov Random Field for High Spatial Resolution Imagery. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6002005. [CrossRef]
6. Wu, J.; Li, B.; Qin, Y.; Ni, W.; Zhang, H. An Object-Based Graph Model for Unsupervised Change Detection in High Resolution Remote Sensing Images. *Int. J. Remote Sens.* **2021**, *42*, 6209–6227. [CrossRef]
7. Fung, T. An Assessment of TM Imagery for Land-Cover Change Detection. *IEEE Trans. Geosci. Remote Sens.* **1990**, *28*, 681–684. [CrossRef]
8. Thonfeld, F.; Feilhauer, H.; Braun, M.; Menz, G. Robust Change Vector Analysis (RCVA) for Multi-Sensor Very High Resolution Optical Satellite Data. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *50*, 131–140. [CrossRef]
9. Wu, C.; Du, B.; Zhang, L. Slow Feature Analysis for Change Detection in Multispectral Imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 2858–2874. [CrossRef]
10. Wang, M.; Han, Z.; Yang, P.; Zhu, B.; Hao, M.; Fan, J.; Ye, Y. Exploiting Neighbourhood Structural Features for Change Detection. *Remote Sens. Lett.* **2023**, *14*, 346–356. [CrossRef]
11. Celik, T. Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and K-Means Clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [CrossRef]
12. Chen, H.; Wu, C.; Du, B.; Zhang, L.; Wang, L. Change Detection in Multisource VHR Images via Deep Siamese Convolutional Multiple-Layers Recurrent Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2848–2864. [CrossRef]
13. Zhu, B.; Yang, C.; Dai, J.; Fan, J.; Qin, Y.; Ye, Y. R_2FD_2: Fast and Robust Matching of Multimodal Remote Sensing Images via Repeatable Feature Detector and Rotation-Invariant Feature Descriptor. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5606115. [CrossRef]
14. Ye, Y.; Zhu, B.; Tang, T.; Yang, C.; Xu, Q.; Zhang, G. A Robust Multimodal Remote Sensing Image Registration Method and System Using Steerable Filters with First- and Second-Order Gradients. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 331–350. [CrossRef]

15. Ye, Y.; Tang, T.; Zhu, B.; Yang, C.; Li, B.; Hao, S. A Multiscale Framework with Unsupervised Learning for Remote Sensing Image Registration. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622215. [CrossRef]
16. Tao, H.; Duan, Q.; An, J. An Adaptive Interference Removal Framework for Video Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5148–5159. [CrossRef]
17. Ye, Y.; Wang, M.; Zhou, L.; Lei, G.; Fan, J.; Qin, Y. Adjacent-Level Feature Cross-Fusion With 3-D CNN for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5618214. [CrossRef]
18. Zhou, Y.; Feng, Y.; Huo, S.; Li, X. Joint Frequency-Spatial Domain Network for Remote Sensing Optical Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5627114. [CrossRef]
19. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SUMLP: A Siamese U-Shaped MLP-Based Network for Change Detection. *Appl. Soft Comput.* **2022**, *131*, 109766. [CrossRef]
20. 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]
21. Hu, Q.; Wang, D.; Yang, C. PPG-Based Blood Pressure Estimation Can Benefit from Scalable Multi-Scale Fusion Neural Networks and Multi-Task Learning. *Biomed. Signal Process. Control* **2022**, *78*, 103891. [CrossRef]
22. Xiang, X.; Tian, D.; Lv, N.; Yan, Q. FCDNet: A Change Detection Network Based on Full-Scale Skip Connections and Coordinate Attention. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6511605. [CrossRef]
23. Li, J.; Hu, M.; Wu, C. Multiscale Change Detection Network Based on Channel Attention and Fully Convolutional BiLSTM for Medium-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 9735–9748. [CrossRef]
24. Jiang, K.; Zhang, W.; Liu, J.; Liu, F.; Xiao, L. Joint Variation Learning of Fusion and Difference Features for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4709918. [CrossRef]
25. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 24 May 2019; pp. 6105–6114.
26. Wen, D.; Huang, X.; Bovolo, F.; Li, J.; Ke, X.; Zhang, A.; Benediktsson, J.A. Change Detection From Very-High-Spatial-Resolution Optical Remote Sensing Images: Methods, Applications, and Future Directions. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 68–101. [CrossRef]
27. Sun, S.; Mu, L.; Wang, L.; Liu, P. L-UNet: An LSTM Network for Remote Sensing Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8004505. [CrossRef]
28. 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]
29. Dai, Y.; Zhao, K.; Shen, L.; Liu, S.; Yan, X.; Li, Z. A Siamese Network Combining Multiscale Joint Supervision and Improved Consistency Regularization for Weakly Supervised Building Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 4963–4982. [CrossRef]
30. Ye, Y.; Zhou, L.; Zhu, B.; Yang, C.; Sun, M.; Fan, J.; Fu, Z. Feature Decomposition-Optimization-Reorganization Network for Building Change Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 722. [CrossRef]
31. Wang, M.; Zhu, B.; Zhang, J.; Fan, J.; Ye, Y. A Lightweight Change Detection Network Based on Feature Interleaved Fusion and Bistage Decoding. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2557–2569. [CrossRef]
32. Li, Z.; Tang, C.; Liu, X.; Zhang, W.; Dou, J.; Wang, L.; Zomaya, A.Y. Lightweight Remote Sensing Change Detection with Progressive Feature Aggregation and Supervised Attention. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602812. [CrossRef]
33. Zhou, F.; Xu, C.; Hang, R.; Zhang, R.; Liu, Q. Mining Joint Intraimage and Interimage Context for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4403712. [CrossRef]
34. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. In Proceedings of the IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210.
35. Song, X.; Hua, Z.; Li, J. PSTNet: Progressive Sampling Transformer Network for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8442–8455. [CrossRef]
36. Fang, S.; Li, K.; Li, Z. Changer: Feature Interaction Is What You Need for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610111. [CrossRef]
37. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224713. [CrossRef]
38. Song, L.; Xia, M.; Weng, L.; Lin, H.; Qian, M.; Chen, B. Axial Cross Attention Meets CNN: Bibranch Fusion Network for Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 32–43. [CrossRef]
39. Song, X.; Hua, Z.; Li, J. LHDACT: Lightweight Hybrid Dual Attention CNN and Transformer Network for Remote Sensing Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 7506005. [CrossRef]
40. Feng, Y.; Xu, H.; Jiang, J.; Liu, H.; Zheng, J. ICIF-Net: Intra-Scale Cross-Interaction and Inter-Scale Feature Fusion Network for Bitemporal Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4410213. [CrossRef]
41. Song, X.; Hua, Z.; Li, J. Remote Sensing Image Change Detection Transformer Network Based on Dual-Feature Mixed Attention. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5920416. [CrossRef]
42. Chu, S.; Li, P.; Xia, M.; Lin, H.; Qian, M.; Zhang, Y. DBFGAN: Dual Branch Feature Guided Aggregation Network for Remote Sensing Image. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *116*, 103141. [CrossRef]

43. Tang, X.; Zhang, T.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. WNet: W-Shaped Hierarchical Network for Remote-Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615814. [CrossRef]
44. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. ISBN 978-3-319-24573-7.
45. Liang, S.; Hua, Z.; Li, J. Enhanced Self-Attention Network for Remote Sensing Building Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 4900–4915. [CrossRef]
46. Chouhan, A.; Sur, A.; Chutia, D. DRMNet: Difference Image Reconstruction Enhanced Multiresolution Network for Optical Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4014–4026. [CrossRef]
47. Zhu, Q.; Guo, X.; Li, Z.; Li, D. A Review of Multi-Class Change Detection for Satellite Remote Sensing Imagery. *Geo Spat. Inf. Sci.* **2022**, 1–15. [CrossRef]
48. Zhang, X.; Yue, Y.; Gao, W.; Yun, S.; Su, Q.; Yin, H.; Zhang, Y. DifUnet++: A Satellite Images Change Detection Network Based on Unet++ and Differential Pyramid. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8006605. [CrossRef]
49. Zhu, S.; Song, Y.; Zhang, Y.; Zhang, Y. ECFNet: A Siamese Network with Fewer FPs and Fewer FNs for Change Detection of Remote-Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6001005. [CrossRef]
50. Cheng, Y.; Cai, R.; Li, Z.; Zhao, X.; Huang, K. Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1475–1483.
51. Saha, S.; Shahzad, M.; Ebel, P.; Zhu, X.X. Supervised Change Detection Using Prechange Optical-SAR and Postchange SAR Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8170–8178. [CrossRef]
52. Fu, Z.; Li, J.; Chen, Z.; Ren, L.; Hua, Z. DAFT: Differential Feature Extraction Network Based on Adaptive Frequency Transformer for Remote Sensing Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5061–5076. [CrossRef]
53. Barkur, R.; Suresh, D.; Lal, S.; Reddy, C.S.; Diwakar, P.G. RSCDNet: A Robust Deep Learning Architecture for Change Detection from Bi-Temporal High Resolution Remote Sensing Images. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 537–551. [CrossRef]
54. 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]
55. 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]
56. 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]
57. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
58. 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]
59. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
60. Yang, H.; Chen, Y.; Wu, W.; Pu, S.; Wu, X.; Wan, Q.; Dong, W. A Lightweight Siamese Neural Network for Building Change Detection Using Remote Sensing Images. *Remote Sens.* **2023**, *15*, 928. [CrossRef]