remote sensing

MDPI

Article

Multi-Scale Fusion Siamese Network Based on Three-Branch Attention Mechanism for High-Resolution Remote Sensing Image Change Detection

Yan Li [1], Liguo Weng [1,*], Min Xia [1], Kai Hu [1] and Haifeng Lin [2]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, B-DAT, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212220011@nuist.edu.cn (Y.L.); xiamin@nuist.edu.cn (M.X.); 001600@nuist.edu.cn (K.H.)
[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; haifeng.lin@njfu.edu.cn
* Correspondence: 002311@nuist.edu.cn

Abstract: Remote sensing image change detection (CD) is an important means in remote sensing data analysis tasks, which can help us understand the surface changes in high-resolution (HR) remote sensing images. Traditional pixel-based and object-based methods are only suitable for low- and medium-resolution images, and are still challenging for complex texture features and detailed image detail processing in HR images. At present, the method based on deep learning has problems such as inconsistent fusion and difficult model training in the combination of the difference feature information of the deep and shallow layers and the attention mechanism, which leads to errors in the distinction between the changing region and the invariant region, edge detection and small target detection. In order to solve the above problems of inconsistent fusions of feature information aggregation and attention mechanisms, and indistinguishable change areas, we propose a multi-scale feature fusion Siamese network based on attention mechanism (ABMFNet). To tackle the issues of inconsistent fusion and alignment difficulties when integrating multi-scale fusion and attention mechanisms, we introduce the attention-based multi-scale feature fusion module (AMFFM). This module not only addresses insufficient feature fusion and connection between different-scale feature layers, but also enables the model to automatically learn and prioritize important features or regions in the image. Additionally, we design the cross-scale fusion module (CFM) and the difference feature enhancement pyramid structure (DEFPN) to assist the AMFFM module in integrating differential information effectively. These modules bridge the spatial disparity between low-level and high-level features, ensuring efficient connection and fusion of spatial difference information. Furthermore, we enhance the representation and inference speed of the feature pyramid by incorporating a feature enhancement module (FEM) into DEFPN. Finally, the BICD dataset proposed by the laboratory and public datasets LEVIR-CD and BCDD are compared and tested. We use F1 score and MIoU values as evaluation metrics. For AMBMFNet, the F1 scores on the three datasets are 77.69%, 81.57%, and 77.91%, respectively, while the MIoU values are 84.65%, 85.84%, and 84.54%, respectively. The experimental results show that ABMFNet has better effectiveness and robustness.

Keywords: remote sensing image change detection; multi-branch attention mechanism; feature extraction; multi-scale feature fusion

Citation: Li, Y.; Weng, L.; Xia, M.; Hu, K.; Lin, H. Multi-Scale Fusion Siamese Network Based on Three-Branch Attention Mechanism for High-Resolution Remote Sensing Image Change Detection. *Remote Sens.* **2024**, *16*, 1665. https://doi.org/10.3390/rs16101665

Academic Editor: Melanie Vanderhoof

Received: 2 April 2024
Revised: 27 April 2024
Accepted: 30 April 2024
Published: 8 May 2024

1. Introduction

Remote sensing CD has always been an important technology in image processing related to remote sensing. It detects two remote sensing photos at different times in the same area, identifies the pixels with semantic changes between them, and finally presents the detection results in the form of black and white images.

The ability and means of obtaining picture data have improved, as evidenced by the emergence of satellite remote sensing technology, the steady maturation of remote sensing imaging technology, and other developments. A large number of high-resolution (HR) remote sensing images need to be processed. HR remote sensing photos provide better geometric and spatial information when compared to medium-resolution and low-resolution images. This makes it easier for people to monitor surface changes with greater accuracy. Remote sensing CD has also become a rising star in the field of remote sensing images, but the results of CD will be disturbed by many factors [1]. On the one hand, there is the influence of remote sensing system, on the other hand, there is the interference of environmental factors, such as the movement, deformation, occlusion of the detected object, or the recognition error caused by camera movement, light change and seasonal change. How to resist the influence of these interference factors and detect the actual changes is very challenging for remote sensing CD tasks. In numerous domains, such as land management [2,3], urban development planning [4,5], environmental monitoring [6,7], disaster prediction [8], and others, remote sensing CD has been extensively employed. In recent years, the research on remote sensing CD has received more and more attention.

So far, different authors have proposed many CD techniques from different perspectives. In general, these technologies can be divided into two categories: traditional technologies and deep learning-based technologies.

The traditional remote sensing CD technology is mainly divided into the pixel-based method and object-based method. The pixel-based method refers to the analysis of each pixel to determine its changes at different time points. The classical method is the difference map method. It obtains the difference map by subtracting the remote sensing images at two time points, and then determines the change area according to the threshold or other methods [9]. The ratio map method divides the remote sensing images at two time points to obtain the ratio map, and then determines the change area according to the threshold or other methods [10]. With the regression analysis method, the basic idea is to obtain the change of image pixels by regression analysis of two remote sensing image data, so as to detect the change of the surface [11]. These methods are simple in theory and easy to quickly identify the change area. However, due to the difficulty in determining the change threshold, the difficulty in extracting the change properties of the target area and the limitation of the surface detection ability, these methods usually cannot obtain complete change information. In order to make better use of the spectral information of the image, multivariate change detection (MAD) [12], iterative reweighted multivariate change detection (IRMAD) [13], principal component analysis (PCA) based on k-means clustering [14], and change vector analysis (CVA) as proposed by Malila [15], came into being. Most of these methods are unsupervised and have achieved good results in ground remote sensing CD tasks. The essence of MAD is canonical correlation analysis (CCA) in multivariate statistical analysis, but this algorithm cannot deal with multi-element remote sensing images better. Therefore, the IR-MAD algorithm is studied and proposed. The core idea of the algorithm is to set the initial weight of each pixel to 1, and assign new weights to the pixels in the two images by iteration. Phase angle data are used by the CVA technique to partition the changes. However, the stability of the algorithm cannot be guaranteed, and the effectiveness of this approach mostly relies on the caliber of the spectral bands used in the computation. Another common algorithm is principal component analysis (PCA) [16]. PCA transforms the image into a set of linear independent representations through linear transformation, which can be used to extract the main feature components of the data. However, PCA depends on the statistical characteristics of the image, so whether the data in the changing region and the invariant region are balanced will have a great impact on the performance of the model. Based on this, Celik [17] proposed an unsupervised CD method that integrates PCA and K-means clustering. However, the pixel-level method is sensitive to high-frequency information in high-resolution remote sensing images, and is easily affected by image geometric correction and radiometric correction errors, and its applicability is limited. Therefore, it is mainly suitable for low- and medium-resolution images. Therefore,

the object-based method is often used in HR remote sensing image CD. It allows richer information representation. The ground features or surface features in the image are used as objects, and their shapes, sizes, and positions at different time points are compared to determine the change. Ma et al. [18] used different segmentation strategies and a series of segmentation scale parameters to test four common unsupervised CD algorithms on urban area images, and studied the influence of several factors on the object-based CD method. Subsequently, by merging multi-scale uncertainty analysis, Zhang et al. [19] presented an object-based unsupervised CD technique. Zhang et al. [20] proposed a method based on the law of cosines with a box–whisker plot to solve the problem that the sample data does not obey the Gaussian distribution, which is superior to the traditional CD method. To some extent, traditional methods that rely on pixels and objects have aided in the development of remote sensing CD. For high-resolution bitemporal images, which contain complex texture features and detailed image detail processing, it is still challenging. Image processing for remote sensing requires new CD approaches.

Natural language processing [21], audio recognition [22,23], and image processing [24–27] have all made extensive use of deep learning techniques in recent years. In order to extract features, the deep learning method does not require the manual creation of feature elements, and it has good learning capabilities [28,29]. Researchers' interest in deep learning-based remote sensing image CDs has grown rapidly as a result of deep learning's achievements in the field of image processing. Convolutional neural networks (CNNs) have been the subject of some outstanding research in the field of remote sensing CD [30,31], thanks to the ongoing advancements in technology. UNet [32], fully convolutional network (FCN) [33], and ResNet [34] structures are commonly employed in the remote sensing CD domain for feature map extraction. In 2016, Gong [35] applied deep neural network to remote sensing image CD for the first time. This method consists of two parts. Firstly, the pre-classification results are obtained by FCM joint clustering, and the training sample set is selected from the pre-classification results. Then, the Restricted Boltzmann Machine (RBM) is used as a tool to construct a deep neural network, and the network parameters are fine-tuned by the back propagation algorithm. Finally, the trained deep neural network outputs the CD results. In 2017, Zhan [36] proposed an optical remote sensing image CD method based on deep twin convolutional neural network.This method transforms the CD problem into an image segmentation problem, and inputs two time images into two deep convolutional neural networks with shared weights to extract image features. Then, the feature distance map is obtained by calculation, and the CD results are obtained directly by clustering or threshold segmentation method. This method effectively improves the high time cost problem brought by the traditional CD method based on deep neural network, and realizes end-to-end remote sensing image CD. With the deepening of research, the remote sensing CD model has been continuously optimized and improved. Wang et al. [37] introduced the attention mechanism into the deep supervised network to capture the link between different scale changes between each module of encoding and decoding to achieve more accurate CD. Yin et al. [38] used the attention fusion module to refine layer by layer in the decoding stage for the reconstruction of the change map, which proved that the introduction of the attention mechanism in the deep learning model can improve the performance and robustness of the model.

The method of image fusion based on deep learning has been widely applied in image processing. The fundamental idea of multi-scale fusion is to utilize features from different scales to complement and enhance each other's information, thereby improving the understanding and description of targets or scenes. Li et al. [39] proposed a novel multi-focus image fusion method using local energy and sparse representation in the Shearlet domain. This method first decomposes the source image into low-frequency and high-frequency subbands through Shearlet transform. Then, it fuses the low-frequency subbands using sparse representation and the high-frequency subbands using local energy. Finally, the fusion image is reconstructed through inverse Shearlet transform. Zhang et al. [40] introduced a Dimension-Driven Multi-Path Attention Residual Block (DDMARB) to effectively capture

multi-scale features. Through the Channel Attention (CA) mechanism, these features are processed differently to better express the depth features of the data. Zhang et al. [41] introduced a compact structure feature distillation block (FD-Block) for high-dimensional information in the encoder stage. The FD module adopts a multi-path feature extraction method, which can fully utilize features from different levels in the high-dimensional input. Based on this, we propose to integrate multi-scale fusion with attention mechanism to fully utilize multi-scale information in images. This not only enhances the understanding and description of image content but also preserves the sharpness of all images in the fused image. However, due to the full integration of different scales of different features and the combination of attention mechanisms, a series of problems such as inconsistent feature fusion, alignment problems, and model complexity and training difficulties will be brought about. The current deep learning algorithm cannot achieve good results in the use of these two, so it cannot guarantee the accuracy of distinguishing between changing regions and invariant regions and the problem of false detection and missed detection of small targets and edge detection. Based on this, we propose a multi-scale feature fusion Siamese network based on an attention mechanism (ABMFNet) in high-resolution remote sensing image CD tasks to solve the above problems. This network contains three modules to assist the network in training. In order to realize the connection between the low-level features and high-level features of spatial difference information, we propose a cross-scale fusion module (CFM) and a bottom-up difference enhancement feature pyramid structure (DEFPN). These two modules are used to fuse the output of different levels of feature difference information to ensure the recovery of the original image detail features during the subsequent upsampling process. In the process of fusing the difference information of the adjacent two layers, we add the feature enhancement module (FEM). This module enhances the representation of the feature pyramid and accelerates the inference speed, while achieving the most advanced performance. In order to solve the problem of insufficient feature fusion and feature connection caused by feature layers of different scales in the fusion process, we designed an attention-based multi-scale feature fusion module (AMFFM). The main contributions of this paper are as follows:

1. To address a series of challenges in change detection tasks for dual-temporal high-resolution remote sensing images, including handling complex texture features, intricate image details, and effectively integrating differing scale features with attention mechanisms, while mitigating issues such as inconsistent fusion, alignment problems, model complexity, and training difficulties, we propose ABMFNet. This method is end-to-end trainable and makes full use of the rich difference information in high-resolution remote sensing images. It can better improve the detection of edges and small targets, and provides a more effective solution for CD in HR remote sensing images.
2. We propose the Attention-based Multi-scale Feature Fusion Module (AMFFM), which not only addresses the issue of insufficient feature fusion and connection between different-scale feature layers, but also enables the model to automatically learn and select important features or regions in the image, concentrating more attention on these areas. Additionally, to better assist the AMFFM module in integrating differential information based on the attention mechanism, we design the Cross-scale Fusion Module (CFM) and the Difference Feature Enhancement Pyramid Structure (DEFPN), facilitating the connection of spatial differential information between low-level and high-level features. Furthermore, the Feature Enhancement Module (FEM) is incorporated into DEFPN to enhance the representation and inference speed of the pyramid.
3. We tested on three datasets; one is the dataset BICD proposed by our laboratory, and the other two public benchmark datasets LEVIR-CD and BCDD. The experimental results show that our proposed model has better improvement and resolution accuracy than the previous algorithms for HR remote sensing image CD.

The rest of the article is divided into the following: The Section 2 introduces the detailed composition and selection advantages of each module of the entire network.

Section 3 introduces three datasets, the ablation experiment on the BICD dataset and proves the performance of the model through experiments. Section 4 discusses the effectiveness of the proposed modules and future prospects. Section 5 is the final summary.

2. Methodology

With the wide attention of remote sensing image CD in the field of image processing, many high-quality algorithms based on CNNs have been applied to this task in recent years [42,43]. The task of remote sensing image CD is essentially a two-class semantic segmentation process, and CNNs perfectly meet this task. Furthermore, CNNs possess advantages such as automatic feature extraction and shared convolutional kernels. These characteristics enable CNNs to effectively capture local patterns and abstract features within images, thereby demonstrating outstanding performance in tasks such as image classification [44]. Therefore, in the coding stage of the network, we select ResNet for feature extraction, and then perform a simple element subtraction operation on the extracted features. The two modules of CFM and DEFPN are used to fuse the difference feature information of different sizes at different stages for subsequent cross-scale fusion. Next, AMFFM is introduced in detail for cross-scale fusion, and the attention mechanism is used to pay more attention to the changing region, while suppressing the process of the invariant region. Finally, the final prediction map is obtained by upsampling. Figure 1 shows the whole network structure.

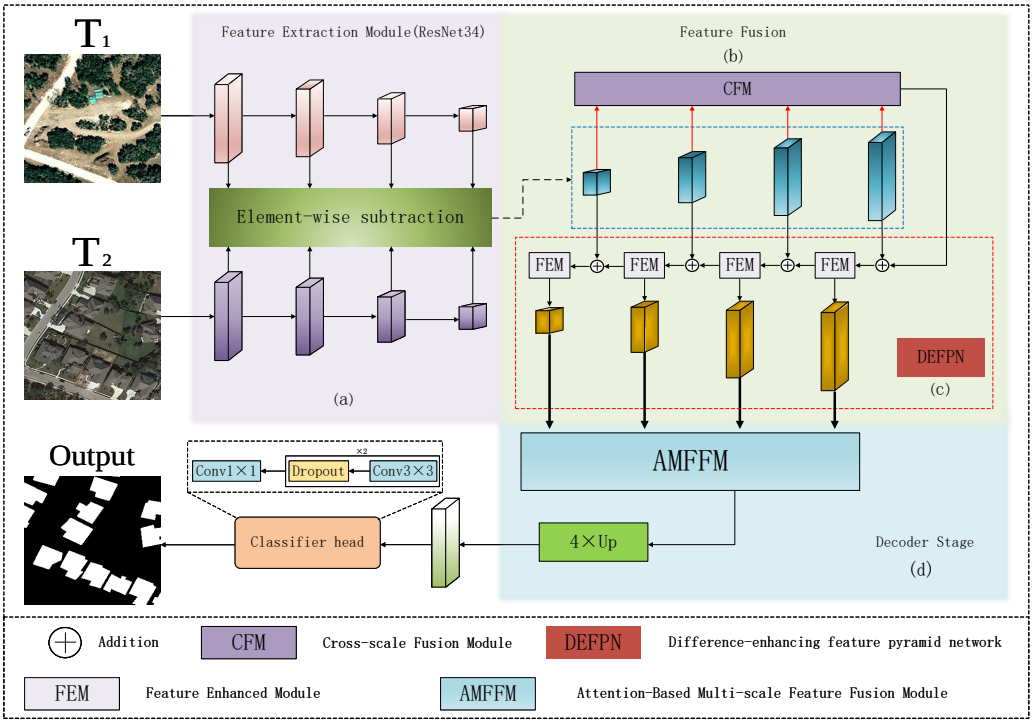

Figure 1. Attention-based multi-scale feature fusion Siamese network (ABMFNet) framework. (**a**) is the feature extraction stage, the backbone network is ResNet34, (**b**) is the feature fusion module, including cross-scale fusion module (CFM) and (**c**) difference enhancement feature pyramid structure (DEFPN), (**d**) is the decoding stage, mainly is the attention-based multi-scale feature fusion module (AMFFM).

2.1. Backbone

The task of the main network is to extract the image feature information of the bitemporal remote sensing image pairs. In the image classification task, the size of the receptive field has an important influence on the classification effect. In order to expand the receptive field, some methods use spatial pyramid pooling [45]. However, the disadvantage of this method is that the calculation speed is slow and the memory consumption is large. Consequently,

as backbone networks, we take into consideration some models that perform better in image classification tasks, like the visual geometry group network (VGG) [46], the densely connected convolutional network (DenseNet) [47], the residual network (ResNet) [34], and the deep learning of deep separable convolution (Xception) [48]. In the CD task, due to the high resolution of the image, the shallow network is generally not used, Thus, the shallow network is unable to adequately extract the image's feature information. The general idea is to design a network as deep as possible. However, as the number of network layers increases, deeper redundant layers will be generated, and it is difficult to train the network. In addition, as the depth of a neural network increases, the problem of vanishing or exploding gradients may arise. However, some network architectures such as ResNet have addressed this issue to a certain extent. Nevertheless, overly deep network structures may increase the risk of overfitting rather than solely reducing the effectiveness of training due to gradient vanishing. Therefore, when designing a network, it is essential to strike a balance between network depth and model complexity, avoiding overly deep structures. Through experimental comparisons, we choose ResNet34 as the backbone network, because ResNet34 effectively tackles the problem of gradient vanishing through residual modeling. The advantage of the residual model is that it can better adapt to the classification function to obtain higher classification accuracy. This structure makes the network very deep, and the training performance is still very good. The core idea is jump connection. As shown in Figure 1a, ResNet generates five feature maps from shallow to deep in the process of feature extraction, but we only use the last four layers, which are 4, 8, 16, 32 times downsampling, and 64, 128, 256, 512 channels of shallow and deep feature maps. The Siamese structure is used to extract the features of the bitemporal remote sensing images T_1, T_2 at the same time, and the feature maps of different sizes $T_{1i}, T_{2i}, i = \{1, 2, 3, 4\}$ are obtained.

2.2. Difference Enhances Feature Pyramid Networks

For the extraction of difference features, we directly use the element subtraction operation after backbone. The four different scale feature maps obtained directly by this operation contain shallow and deep difference information, respectively. However, if we directly use the information from these four difference features for subsequent upsampling prediction, it will lead to a smaller size of the deep network and relatively less geometric information. At the same time, the shallow information contains few image semantic features, which is not conducive to CD. So, after the backbone, we incorporated a feature fusion step. In computer vision tasks, feature pyramids [49] are usually used to fuse feature maps with different resolutions, reflecting certain advantages. Based on this, this paper proposes a bottom-up difference enhancement feature pyramid structure for encoding.

The structure of this module is shown in Figure 1c. Figure 2a shows the operation of one layer; we fuse the difference feature maps of different sizes by pooling convolution for down-sampling and element addition. The feature enhancement module is added to the fusion process of the difference information of the adjacent two layers, which enhances the inference speed of the feature pyramid and improves the performance of the model. Figure 2b shows the structure diagram of the feature enhancement module. The difference feature size obtained by subtracting the elements is $f_{\text{in}} \in \mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$; we adopt a two-branch structure, where one is responsible for extracting channel change information, and the other is responsible for guiding the original features. The upper branch obtains the useful channel information of the differential features through the Global Average Pooling layer, and the obtained feature map is $\mathrm{C} \times 1 \times 1$, which is then multiplied by the original feature information of the lower branch to restore the feature size, and then the elements are added to the original features and, finally, the output features $f_{\text{out}} \in \mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ are obtained through the 1×1 convolution layer. The calculation formula of the above process is:

$$f_{\text{out}} = RELU\left(B\left(f^{1\times1}(f_{\text{in}} + \sigma(f_{\text{in}} \otimes GAP(f_{\text{in}})))\right)\right) \quad (1)$$

In this formula, $RELU(\bullet)$ represents the $RELU$ activation function, $B(\bullet)$ indicates batch standardization processing, $f^{1\times1}$ denotes two-dimensional convolution with convolu-

tion kernel of 1, $\sigma(\bullet)$ is the *Sigmoid* activation function, $\otimes$ represents element multiplication, and $GAP(\bullet)$ is global average pooling.

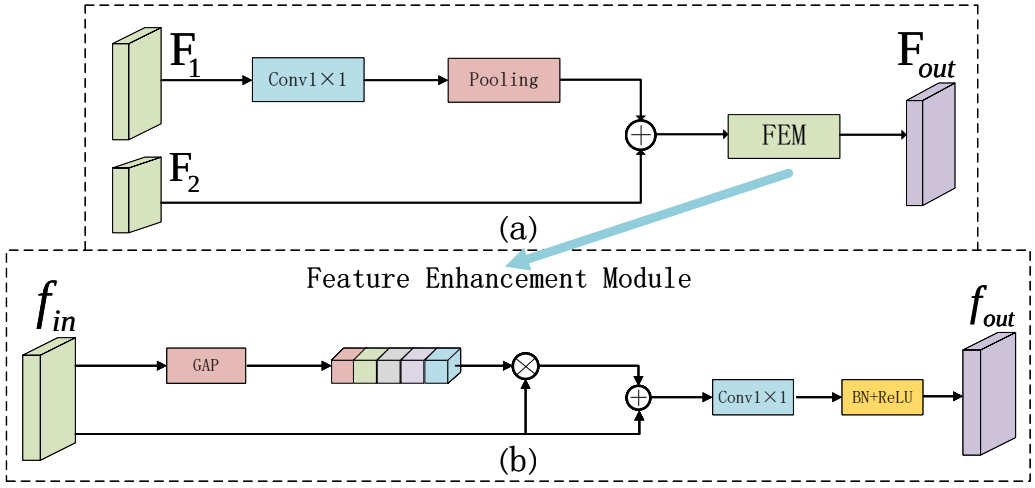

Figure 2. Structure of the difference enhancement feature pyramid. (**a**) shows a layer of display of the difference enhancement feature pyramid module, and (**b**) shows a feature enhancement module.

The difference features of multiple scales obtained by element subtraction are f_i, $i = \{1, 2, 3, 4\}$; firstly, it is fused with the previous layer of difference features adopted under feature enhancement and pooling. On the one hand, the feature enhancement features are obtained as output f_i', $i = \{1, 2, 3, 4\}$ and, on the other hand, they are transmitted to the next layer as input through the same operation. The calculation formula of the difference enhancement feature pyramid is as follows:

$$f_1' = F(f_1) \tag{2}$$

$$f_i' = F(f_i + AvgPool(f_{i-1}')), i = \{2, 3, 4\} \tag{3}$$

In this formula, $F(\bullet)$ represents FEM, $AVGPool(\bullet)$ represents a 2-fold average pooling downsampling layer. The difference characteristic formula is as follows:

$$f_i = abs(T_{1i} - T_{2i}), i = \{1, 2, 3, 4\} \tag{4}$$

Here, $abs(\bullet)$ is element subtraction.

2.3. Cross-Scale Fusion Module

In the process of DEFPN module's fusion of different features at different stages, the output of shallow feature information is not accurately modeled due to the relationship with deep features, resulting in that the subsequent feature fusion related to each semantic class does not reach the best state. Currently, there are many cross-scale fusion architectures. The CSFF module proposed by Chen et al. [50] integrates contextual information of features to better achieve target feature extraction, addressing the issue of inter-class similarity. Inspired by this, we design a cross-scale fusion module, which captures the context dependencies of different stages through a guiding mechanism, and adds semantic and detailed information to shallow features to generate richer representations. The structure is shown in Figure 3.

Multiple scale feature maps obtained by element subtraction are $f_i, i = \{1, 2, 3, 4\}$; since the features of each stage have different channel numbers and resolutions, we first use the 1 × 1 convolution layer to sample them up to the same size 64 × 128 × 128 by bilinear interpolation. The expanded feature map is $f_i', i = \{1, 2, 3, 4\}$; then, the adjacent feature maps f_i' are densely linked to obtain four new features $F_i, i = \{1, 2, 3, 4\}$. Finally, the concat operation is used to stitch on the channel dimension, and then the number

of channels is restored through the 1 × 1 convolutional layer to obtain the final output feature. Therefore, f_i encodes the detailed information from the shallow layer and the deep semantic information jointly and, finally, a feature map with richer context information and stronger representation is obtained. The calculation formula of the above process is as follows:

$$f_i' = Upsample(f^{1\times1}(f_i)), i = \{1,2,3,4\} \quad (5)$$

$$F_1 = f_1' + f_2' \quad (6)$$

$$F_2 = f_1' + f_2' + f_3' \quad (7)$$

$$F_3 = f_2' + f_3' + f_4' \quad (8)$$

$$F_4 = f_3' + f_4' \quad (9)$$

$$f_{out} = f^{1\times1}([F_1; F_2; F_3; F_4]) \quad (10)$$

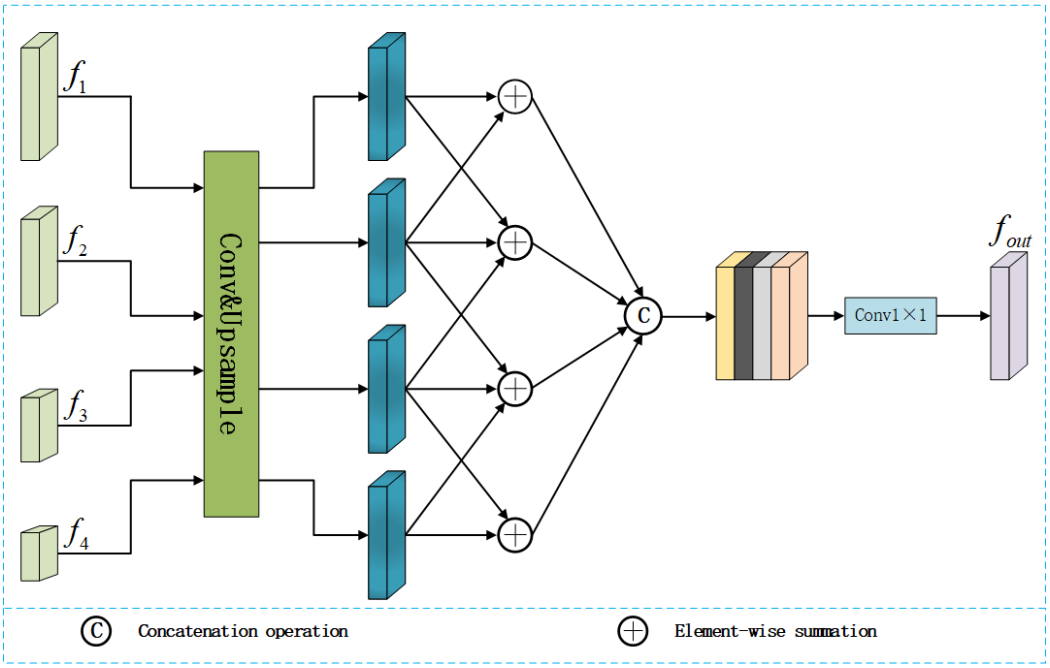

Figure 3. Cross-scale fusion module structure diagram.

In this formula, $Upsample(\bullet)$ represents the bilinear interpolation upsampling, batch normalization and $ReLU$ activation function, and $[\bullet]$ refers to the Concat splicing operation.

2.4. Attention-Based Multi-Scale Feature Fusion Module

In recent years, with the in-depth study of remote sensing CD tasks, many deep learning approaches solely use basic feature extraction networks, which produces unsatisfactory results for CD tasks. On the one hand, simple feature extraction networks are heavily influenced by semantic interferences such as shooting angles, seasonal changes, and lighting variations. They often struggle to accurately label changing regions in scenarios with densely arranged objects or complex shapes. On the other hand, they fail to fully leverage multi-scale information. The fusion of multi-scale features can help establish connections between them, thereby enhancing the performance of our network. There are two common types of multi-scale feature fusion networks: parallel multi-branch networks, as seen in Inception [51], SPPNet [52], DeepLabV2 [53], and PSPNet [54]; and serial skip-connection structures, as used in FPN [49], UNet [32], HRNet [55], PANet [56], and BiFPN [57]. These architectures perform feature extraction at different receptive fields. Although multi-scale fusion adequately considers features at different scales to merge local details and global

structures, they cannot automatically learn and select important features or regions in the image and focus more attention on these areas. Therefore, we introduce attention mechanisms into multi-scale fusion. The edge information of the changing region can be better obtained by using the layer-by-layer fusion of adjacent features, and the attention mechanism [58] can better focus on the changing region and suppress the invariant region. Based on this, we propose an attention-based multi-scale feature fusion module (AMFFM).

Figure 4 describes the detailed content of the entire module. The input is the four different scale difference feature maps $f_i^{'}, i = \{1, 2, 3, 4\}$ obtained by the DEFPN module. The feature maps of the adjacent scales are input into the AFFM for feature fusion. In the fusion process, the resolution of the semantic features in the multi-scale feature map is continuously restored, and the attention to the changing region is enhanced. Therefore, we use 6 AFFMs to obtain a feature map that is restored to the 1/4 size of the input image. In the subsequent use of 4-times upsampling and a classification head composed of convolution to obtain classification results. AFFM is shown in Figure 4b. The feature maps of different stages have different channel numbers and sizes, therefore, a 2-times upsampling operation is performed for deep features, and then spliced with the adjacent upper layer features to obtain the feature map of 2C×H×W. Since the direct cross-level feature fusion method does not take into account the importance and interactivity of the channel and spatial dimension, we send the obtained feature map to the attention module. The feature map is recalibrated by space and channel, which enhances the network's attention to more noteworthy feature information. Based on CAM and SAM in CBAM, we propose a new attention module, whose structure is shown in Figure 4c. Different from CBAM's method of connecting CAM and SAM in series, we adopt the idea of three branches and parallel connection. One branch is used to guide the original features, and the other two are used to extract the change information of the channel and spatial dimension of the original feature map, respectively. The three branches are fused by jump connection to achieve mutual guidance. That is, the original feature map is first refined by CAM and SAM to obtain the channel and space, and then multiplied by the elements of the original feature to restore the size, and then spliced with the original feature. Next, the consistency of the number of channels is maintained by 1 × 1 convolution. Finally, the three branch elements are added to obtain the final output feature map. The formulae of the above process are:

$$f_c = [f_{in}; CAM(f_{in}) \otimes f_{in}] \tag{11}$$

$$f_s = [f_{in}; SAM(f_{in}) \otimes f_{in}] \tag{12}$$

$$f_{out} = f^{1\times1}(f^{1\times1}(f_{in} + f^{1\times1}(f_c) + f^{1\times1}(f_s))) \tag{13}$$

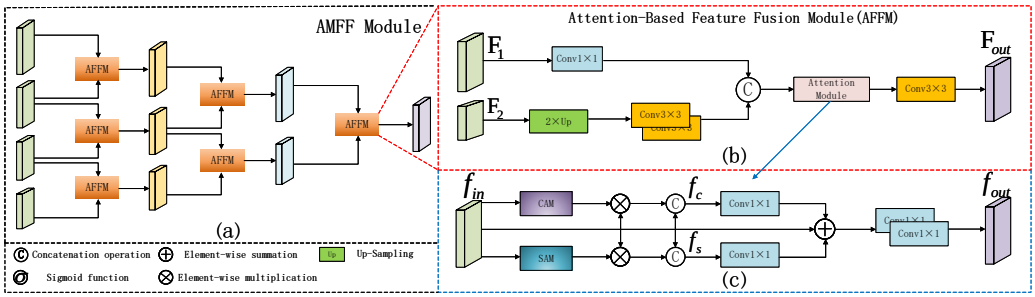

Figure 4. (**a**) shows the Attention-based Multi-scale Fusion Module (AMFFM), (**b**) shows the Attention-based Feature Fusion Module (AFFM), (**c**) shows the details of three-branch attention.

Here, $f^{1\times1}$ represents a two-dimensional convolution with a convolution kernel of 1, batch normalization and *ReLU* activation function, f_{in} represents the input feature map, CAM represents channel attention, SAM represents spatial attention, f_c is the feature map after CAM and f_{in} processing, and f_s is the feature map after SAM and f_{in} processing.

The size of the output feature map is 2C × H × W. Then, we use the 3 × 3 convolution block to keep the number of channels of the feature map consistent. The formula below illustrates the aforementioned computation process:

$$F_{out} = f^{3\times3}(Atten([f^{1\times1}(F_1); f^{3\times3}(Up(F_2))])) \quad (14)$$

In the formula, $f^{3\times3}$ is a 3 × 3 convolution block, $Atten$ represents a new attention module, and Up is a bilinear interpolation 2-times upsampling.

Figure 5 shows the heat map after we use the AFFM module. (a) and (b) are the original images, (c) is the label, (d) is the heat map of the backbone network without the AFFM module, and (e) is the heat map with the AFFM module. Areas where the original attention is not obvious, or there is a large area of focus error, can be seen; that is, the red in the feature map represents a highly concerned area, and the blue and green represent a lower weight area. After the introduction of the AFFM module, the feature map is more effective and accurate for these areas. Next, the two modules of CAM and PAM are introduced.

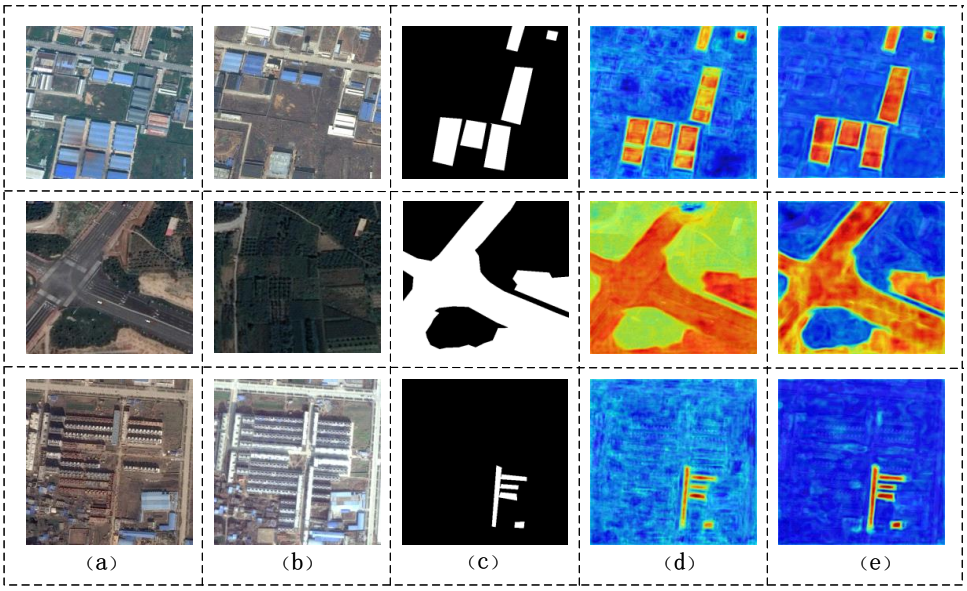

Figure 5. Comparison of heat maps with AFFM module and without AFFM module. (**a**,**b**) original bitemporal image, (**c**) label, (**d**) heat map without AFFM module, (**e**) heat map with AFFM module.

2.4.1. Channel Attention Module

Figure 6 shows the structure of the channel attention module, which keeps the channel dimension unchanged, compresses the spatial dimension, and focuses on the meaningful information in the input features. The size of the input feature map is 2C × H × W. Firstly, the spatial information is squeezed by average pooling and maximum pooling operations to refine the channel, the feature map is compressed into two tensors of 2C × 1 × 1. Then, the two tensors are input into a shared multi-layer perceptron (MLP) with a hidden layer, and the obtained output is combined into a feature vector by element-by-element summation. Finally, the importance of each channel is obtained by assigning different weights to each channel through *sigmoid* function excitation. The formula can be described as:

$$F_{ca} = \sigma(MLP(AvgPool(f_{in})) + MLP(MaxPool(f_{in}))) \quad (15)$$

Here, $\sigma(\bullet)$ represents the *sigmoid* activation function, f_{in} represents the input feature map, and F_{ca} refers to the feature map obtained by CAM.

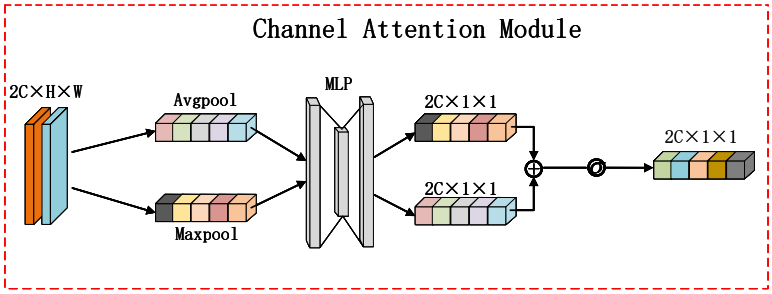

Figure 6. Channel attention module.

2.4.2. Spatial Attention Module

Figure 7 is the spatial attention module structure, which keeps the spatial dimension unchanged, compresses the channel dimension, and focuses on the position information of the target. Its input is the output after multiplying the original feature elements by CAM. Firstly, the channel information is squeezed by average pooling and maximum pooling operations to refine the spatial dimension, and the feature map is compressed into two tensors of 1 × H × W. Then, the Concat splicing operation is performed on two layers, and the feature map of 1 channel is transformed by 7 × 7 convolution. Finally, the feature map output of SAM is obtained by a sigmoid function. The formula is as follows:

$$F_{sa} = \sigma(f^{7\times7}([AvgPool(f_{in}); MaxPool(f_{in})])) \tag{16}$$

Here, $\sigma(\bullet)$ represents the *sigmoid* activation function, and $f^{7\times7}$ represents the two-dimensional convolution with a convolution kernel of 7, batch normalization, and *ReLU* activation function. F_{sa} refers to the feature map obtained by SAM.

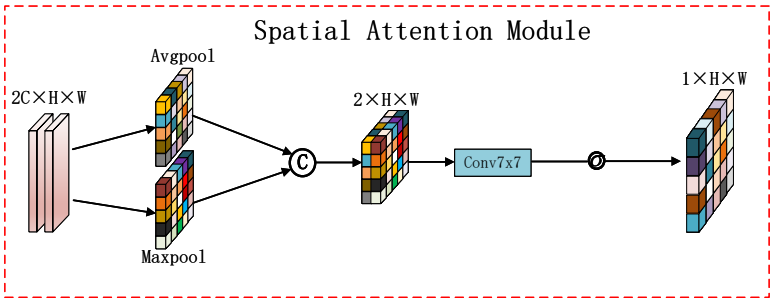

Figure 7. Spatial attention module.

2.5. Loss Function

In the training phase, we use the BCEWithLogitsLoss binary classification loss function. BCEWithLogitsLoss is a combination of binary cross entropy loss function and *sigmoid* function. It not only solves the problem of instability of *sigmoid* function, but also avoids the problem of gradient disappearance in training. Its calculation method is to process the real label and the predicted score by the *sigmoid* function first, and then use the two-class cross entropy to calculate the loss. The formula is as follows:

$$l_i = y_i \cdot \log(\sigma(p_i)) + (1 - y_i) \cdot \log(1 - \sigma(p_i)) \tag{17}$$

$$BCEWithLogitsLoss = \frac{1}{N}\sum_{i=1}^{N}(l_i) \tag{18}$$

This function compares the input-predicted value with the real value, and calculates the difference between the predicted value and the real value. In the formula, $\log(\bullet)$ is the natural logarithm, p_i is the predicted value of the model output, and y_i is the true value of the label. This function helps the model gradually improve the prediction ability and better predict the target.

2.6. Learning Rate and Evaluation Index

We use Adam gradient descent as our network optimizer and BCEWithLogitsLoss as our network loss function. The selection of learning rate in deep learning task is very important. The model will not converge and the loss value will skyrocket if the learning rate is too high. If the learning rate is too low, the complexity of the fusion network will be greatly increased, and the model will be difficult to converge. We have set the attenuation index to 0.9, the batch size to 6, the learning rate to 0.001 at the beginning, and a maximum of 200 training iterations, as per our experiment. The assessment metrics that were employed were joint average intersection (MIoU), precision (PR), recall rate (RC), pixel accuracy (PA) and F1 score (F1). PA represents the proportion of correctly predicted change regions in all pixels, the ratio of successfully predicted change regions in the prediction map to the total number of pixels in all true reference change regions is represented by PR, whereas the proportion of correctly detected change regions in the original picture is represented by RC. The prediction results' total assessment metrics are represented by the F1 score. The better the forecast outcomes, the higher their values. MIoU is to calculate the ratio between the intersection and union of two sets, which represents the change region and the invariant region in the CD task. By combining different evaluation indicators, the performance of the model can be evaluated more comprehensively. The above formulae are as follows:

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

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$F1 = \frac{2 \times PR \times RC}{PR + RC} \tag{22}$$

$$MIoU = \frac{TP}{TP + FP + FN} \tag{23}$$

In the formulae, TP represents true positive and is a correctly predicted change area; FP represents false positive, which is that the invariant region is incorrectly predicted as the change region; TN represents true negative, and is an invariant region for correct prediction; and FN represents false negative, which refers to the wrong prediction of the changing region as the invariant region.

3. Experiment

All the experiments in this section are completed on pytorch and RTX3070Ti. In order to verify the effectiveness and robustness of the model, we designed ablation experiments and comparative experiments; that is, we completed the selection and ablation experiments of the backbone network on the BICD dataset, and completed comparative experiments on the LEVIR-CD dataset and the BCDD dataset. Experiments show that our model is more effective and advanced than other algorithms in CD tasks.

3.1. Dataset

To validate its effectiveness, we conducted tests on our in-house dataset and two publicly available benchmark datasets. In this section, we first analyze the three datasets. By examining the change regions in each image, we can gain a comprehensive understanding of the dataset composition. As shown in Table 1, the proportion of change regions in the three datasets is relatively small compared to unchanged areas. To provide a more intuitive representation, we counted the number of images at different proportions. As illustrated in Figure 8, the horizontal axis represents the proportion of change area to the overall image, while the vertical axis indicates the number of images within each proportion. From

the data analysis graph, it can be observed that the majority of images have change area proportions within 20%.

Table 1. Statistics of the number of changed pixels and the number of invariant pixels in BICD, LEVIR-CD and BCDD datasets.

Dataset		Changed Pixels	Unchanged Pixels	Ratio
BICD	train	33,311,501	613,798,899	0.054
	val/test	21,277,278	128,144,802	0.166
	total	54,588,779	741,943,701	0.074
LEVIR-CD	train	21,412,971	445,203,349	0.048
	val/test	6,837,404	127,380,324	0.054
	total	28,250,375	572,583,673	0.049
BCDD	train	15,422,878	82,029,154	0.188
	val/test	3,751,149	20,562,707	0.182
	total	19,174,027	102,591,861	0.187

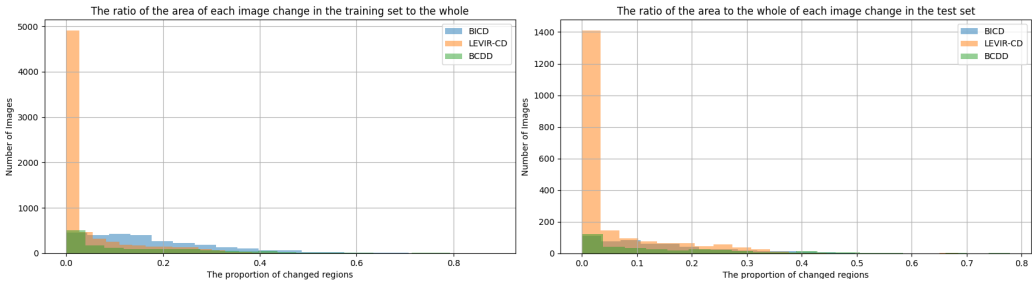

Figure 8. Data analysis. The ratio of the area of change in each image in the dataset to the whole.

3.1.1. BICD Dataset

The BICD dataset contains 3420 pairs of 512 × 512 resolution bitemporal remote sensing images. It is a data set collected by our laboratory from Google Earth to verify the validity of the model. Among them, 2280 were used for training, 570 for verification, and 570 for testing. Compared with other datasets, the remote sensing images in this dataset are composed of photo pairs taken at different locations in eastern China from 2010 to 2019, including various industries, farms, highways and various buildings. The label of the data set is manually completed by us and a unified clipping is performed. In order to avoid the partial change area being cut into small sub-regions during the cutting process, it is ensured that each complete change area has a corresponding slice, and some areas between the slices overlap to ensure that the information of the change area is accurate enough. As shown in Figure 9, we show some representative samples from various categories of the dataset. We can observe that our dataset has more objects, environmental information, and more types of changes. In order to replicate the actual application settings as close as possible, we intentionally include several image pairs with high angle deviations. We also selected some image pairs captured in various seasons to consider seasonal variations. In addition, the image pairs of the unaltered regions are added to the data set to appropriately increase the imbalance between the samples in order to more clearly distinguish the advantages and disadvantages of the model.

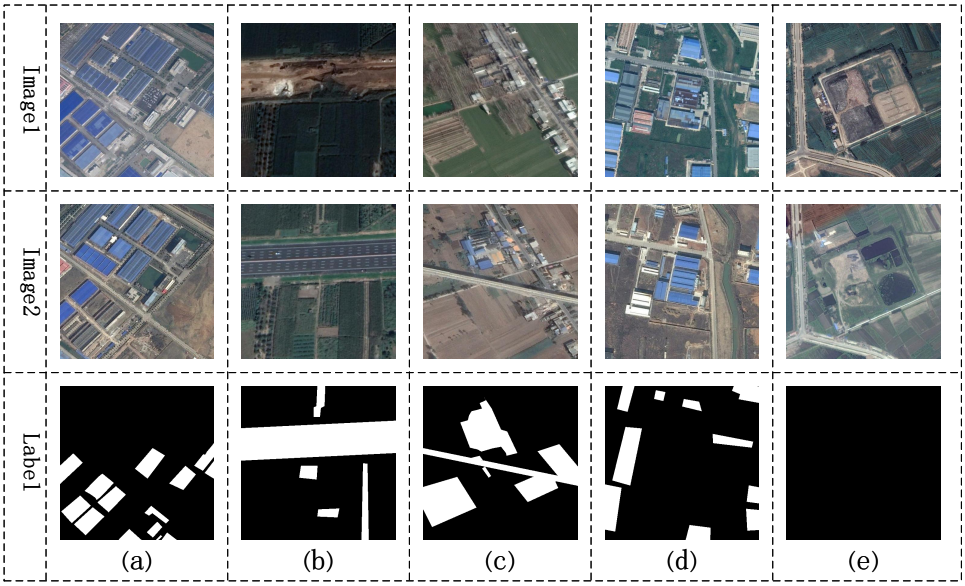

Figure 9. The BICD dataset map shows the changing areas and invariant areas of factories, highways, and farmland. (**a**–**e**) denote factory, farmland, road, building and unchanged area, respectively.

Model training is a complex and time-consuming process that requires a large number of diverse samples to improve its generalization performance. To increase the diversity and quantity of data samples, it is often necessary to perform data augmentation on training images. Data augmentation plays a crucial role in deep learning, as it effectively enhances the model's generalization performance by introducing various transformations and noise. Common image data augmentation methods include rotation, flipping, scaling, cropping, translation, adding noise, adjusting HSV saturation, and sharpening images. These methods enable the model to better understand images, thereby improving its ability to generalize across various scenarios. During the data augmentation process, it is important to pay attention to certain parameter settings to ensure the effectiveness of the augmentation effects. As shown in Table 2, the probability of horizontal or vertical rotation of images is typically set to 50%, and the variation range of HSV saturation is also 50%. Additionally, to avoid introducing too many negative samples, rotational enhancement is generally limited to plus or minus 10 degrees, while translation augmentation should be within 20% of the image's own length. It is worth emphasizing that all these data augmentation methods are only used during the training phase, and no data augmentation should be applied during model testing to ensure the accuracy and reliability of the test results. Ultimately, through these data augmentation methods, the image samples during the training process can be effectively enriched, as shown in Figure 10, thereby improving the model's generalization performance and enabling it to better adapt to different input conditions and scenarios.

Table 2. The parameters of data augmentation.

Method	Parameter
Random Rotation	−10°~10°
Translation	−20°~20°
HSV Saturation	−50°~50°
Random Horizontal Flip	50%
Random Vertical Flip	50%

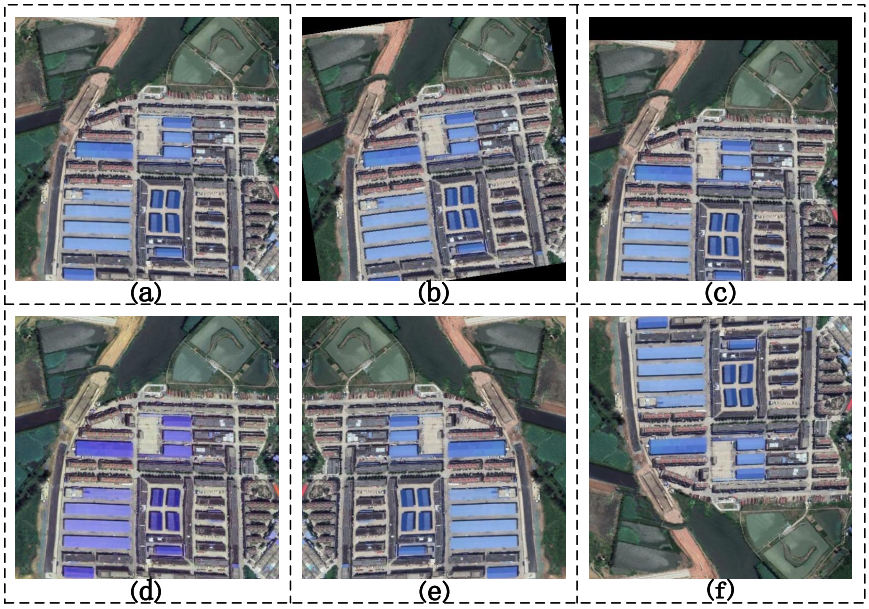

Figure 10. Charts of different data augmentation methods. (**a**) is the original image; (**b**) is the original image rotated by 10 degrees; (**c**) is the translated image; (**d**) is the image after HSV transformation; (**e**) is a horizontally flipped image; (**f**) is a vertically flipped image.

3.1.2. LEVIR-CD Dataset

LEVIR-CD is a massive dataset with 637 pairs of extremely high-resolution (1024 × 1024 pixels) Google Earth pictures that are used to identify changes in buildings. LEVIR-CD focuses on photos taken between five and fourteen years ago that show notable changes in land use, particularly in the area of structures. These images include a variety of building types, including high-rise flats, villas, modest garages, and enormous warehouses. In addition, the effects of light and seasonal variations are particularly pronounced in this dataset. The completely annotated LEVIR-CD has 31,333 unique examples of change construction. As shown in Figure 11, we selected three types of building image display, namely building additions, disappearances, and no changes. Each sample is reduced to 16 small blocks with a size of 256 × 256, generating 7120 image block pairs for training and 2048 for verification and testing.

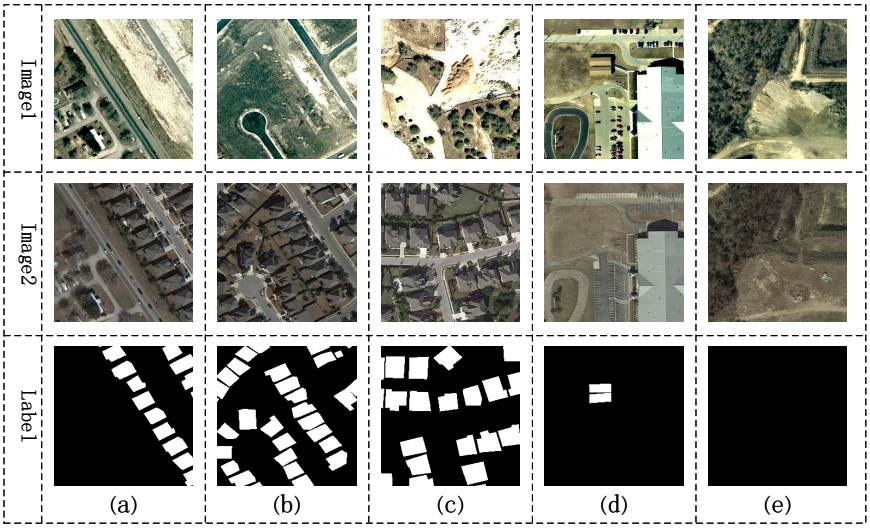

Figure 11. LEVIR-CD dataset. (**a**–**c**) shows the areas of added buildings, (**d**) shows the areas of removed buildings, and (**e**) shows the unchanged areas.

3.1.3. BCDD Dataset

The BCDD dataset was published by Ji et al. [59] from Wuhan University, as shown in Figure 12. It covers the 6.3 magnitude earthquake area in Christchurch, New Zealand, in February 2011, and the subsequent reconstruction area. The dataset contains aerial images obtained in April 2012, including 12,796 buildings of 20.5 square kilometers (16,077 buildings in the same area of the 2016 dataset). By manually selecting 30 GCPs on the ground, the sub-dataset is geographically corrected into an aerial image dataset with an accuracy of 1.6 pixels. Since the size of the image is 32,507 × 15,324, we divide the two images into non-overlapping 256 × 256 pixel image pairs. Then, we select 1487 image pairs in the cropped image as the training set and 371 as the validation set and test set.

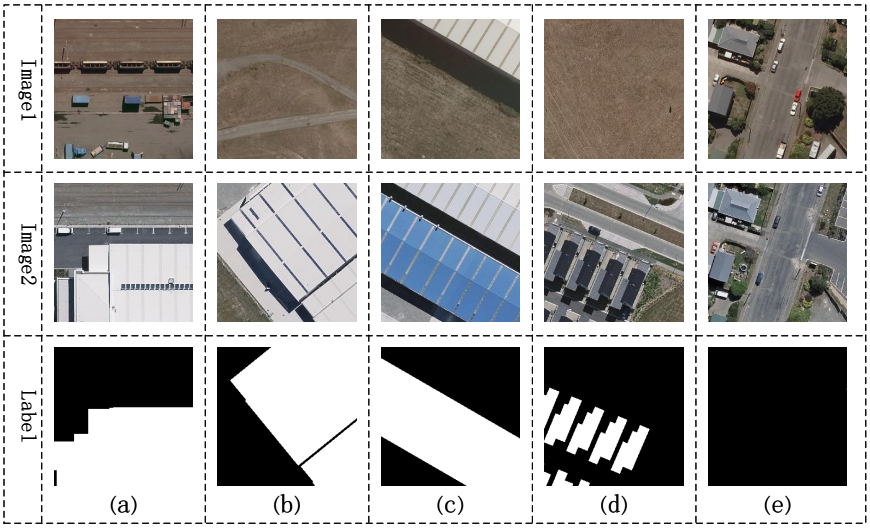

Figure 12. BCDD dataset graph; the first line is the unchanged graph, the second line is the changed graph, and the third line is the labels. (**a**–**d**) shows the changing areas, (**e**) shows the unchanged areas.

3.2. Selection of Backbone Network

With all training parameters set the same, we tested several different backbone networks, as shown in Table 3. We selected VGG_16, VGG_19, DenseNet_121, ResNet_18, ResNet_34, and ResNet_50 as backbone networks for experimentation. The best-performing data are highlighted in bold. It can be observed that ResNet34 performs the best, so we chose ResNet_34 as the backbone network for subsequent experimental testing.

Table 3. Selecting the test results of different backbone networks (bold represents the best result).

Method	PA (%)	RC (%)	PR (%)	F1 (%)	MIoU (%)
VGG_16	91.28	65.01	72.78	66.28	76.11
VGG_19	92.35	67.24	74.05	68.17	78.32
DenseNet_121	94.61	73.42	78.72	72.04	82.68
ResNet_18	95.93	74.51	80.92	75.86	83.56
ResNet_34	**96.15**	**76.36**	**81.81**	**77.69**	**84.65**
ResNet_50	96.01	74.63	81.12	76.01	83.74

3.3. Ablation Experiments on BICD Dataset

In this section, we can better understand and verify whether the proposed module can effectively improve the accuracy of the model by adding and deleting some modules on the ResNet_34 backbone network. Table 4 shows the results of ablation experiments.

Table 4. Ablation experiment of ABMFNet (bold represents the best result).

Method	PA (%)	RC (%)	PR (%)	F1 (%)	MIoU (%)
ResNet34	95.44	73.1	80.2	74.62	82.28
ResNet34 + CBAM	95.67	75.29	79.04	75.41	82.85
ResNet34 + AMFFM	95.92	75.76	81.27	76.64	83.92
ResNet34 + AMFFM + DEFPN	96.09	75.89	81.8	77.26	84.39
ResNet34 + AMFFM + DEFPN + FEM	96.11	76.28	81.33	77.39	84.49
ResNet34 + AMFFM + DEFPN + FEM + CFM	**96.15**	**76.36**	**81.81**	**77.69**	**84.65**

Attention mechanisms can adaptively calibrate changes in channel and spatial dimensions. By utilizing attention fusion modules, the network can better focus on changing regions while suppressing invariant areas, making it one of the important methods to improve the accuracy and effectiveness of the network. Experimental results in Table 4 show that after adding the AMFFM module, the F1 score and MIoU values increased by 2.02% and 1.64%, respectively. Furthermore, compared to the CBAM module, it can be observed that the network with the AMFFM module performs better, with the F1 score and MIoU increasing by 1.23% and 1.07%, respectively, demonstrating the effectiveness of AMFFM. To establish connections between low-level and high-level features regarding spatial differences, we proposed the DEFPN module to facilitate better hierarchical connections during the encoding stage and fuse the output of feature difference information across different levels. The data in Table 4 indicate that the DEFPN module shows a decent improvement in fusion, with the F1 score and MIoU values increasing by 0.64% and 0.47%, respectively. Adding the FEM module during the fusion process not only accelerates the inference speed of the pyramid, but also effectively improves the model's performance. Experimental results in Table 4 show that the F1 score and MIoU increase by 0.13% and 0.1%, respectively. The CFM module captures contextual dependency relationships across different stages through guidance mechanisms, adding semantic and detail information to shallow features to produce richer representations. The data in Table 4 demonstrate that the proposed CFM module enhances the F1 score and MIoU by 0.3% and 0.16%, respectively.

The ablation experiment clearly shows that the proposed auxiliary module has a significant improvement and optimization on the performance and evaluation indicators of the overall network. When the three modules jointly guide the backbone network, the final network indicators PA, RC, PR, F1 and MIoU are proposed, respectively.

3.4. Model Performance Comparison

In this section, we conduct a comprehensive comparison between our proposed model ABMFNet and a range of existing change detection models, including traditional ones such as FC_EF, FC-Siam-Diff, FC-Siam-Conc [60], SNUNet [61], ChangeNet [62], TCDNet [63], DASNet [64], MFGAN [43], TFI-GR [65], and SAGNet [38], as well as deep learning models like FCN8s [33], SegNet [66], UNet [32], HRNet [55], DeepLabV3+ [67], and BiseNet [68]. We evaluate these models based on several criteria, including their complexity and parameter count, as well as F1 score and MIoU values. To provide a more intuitive representation, we present scatter plots in Figure 13. Our analysis reveals that ABMFNet demonstrates moderate levels of complexity and parameter count. More importantly, compared to both traditional and deep learning-based models, ABMFNet achieves significant improvements in both the F1 score and MIoU values. Furthermore, when compared to recent effective models such as MFGAN, TFI-GR, and SAGNet, ABMFNet outperforms them in terms of F1 score and MIoU values while adding only a minimal amount of FLOPs and reducing the parameter count. This highlights the effectiveness and efficiency of our proposed ABMFNet model in remote sensing change detection tasks.

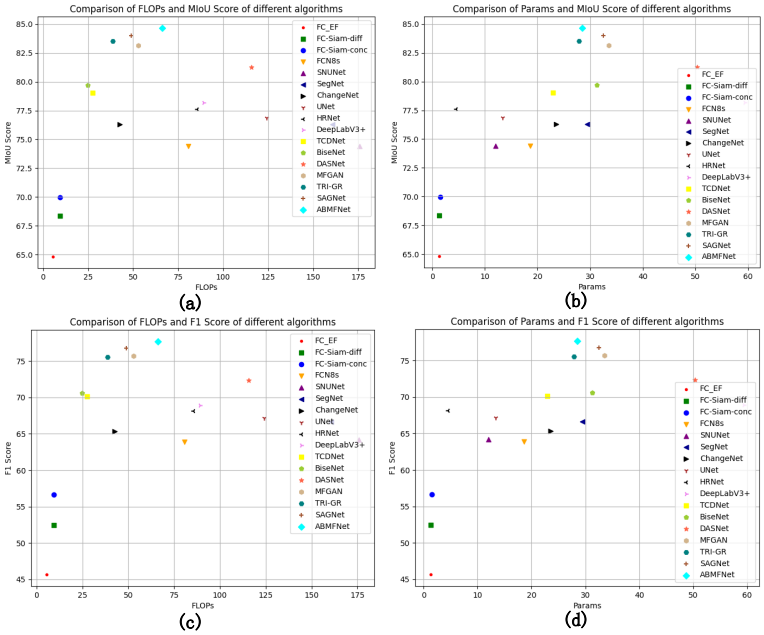

Figure 13. Comparison of the performance of different models. (**a**) shows the model's complexity and MIoU Score, (**b**) shows the model's parameter count and MIoU Score, (**c**) shows the model's complexity and F1 Score, (**d**) shows the model's parameter count and F1 Score.

3.5. Analysis of Experimental Results

3.5.1. Comparative Experiments on BICD Dataset

Table 5 presents the evaluation metrics of various algorithms on the BICD dataset, where all deep learning models use consistent training parameters to ensure the objectivity of comparative experiments. From Table 5, it can be observed that the FC_EF model has the lowest F1 score and MIoU indicator, with only 45.67% and 64.83%, respectively. Considering the characteristics of change detection tasks, we optimized the model. After integrating multi-scale fusion with multi-branch attention mechanisms, the evaluation metrics improved significantly, with PA, RC, PR, F1, and MIoU reaching 96.15%, 76.36%, 81.81%, and 84.65%, respectively. This confirms the effectiveness of ABMFNet.

Table 5. Comparing the experimental results of different algorithms, the input image size of the model is (3 × 512 × 512) (bold numbers represent the optimal results).

Method	PA (%)	RC (%)	PR (%)	F1 (%)	MIoU (%)	FLoPs (G)	Time (s)
FC_EF	90.32	40.4	69.3	45.67	64.83	**5.52**	8.42
FC-Siam-Diff	91.55	46.56	73.30	52.46	68.36	9.29	8.09
FC-Siam-Conc	91.51	53.79	69.21	56.62	69.97	9.34	7.67
FCN8s	92.93	64.44	67.41	63.91	74.41	80.68	4.91
SNUNet	92.98	64.02	69.24	64.22	74.41	175.76	10.24
SegNet	93.75	64.38	72.75	66.62	76.28	160.72	5.99
ChangeNet	94.02	61.08	75.24	65.35	76.29	42.73	15.87
UNet	93.89	65.11	73.53	67.15	76.86	124.21	**4.27**
HRNet	94.10	66.8	74.21	68.18	77.61	85.62	18.34
DeepLabV3+	94.39	67.63	74.62	68.92	78.17	89.15	15.38
TCDNet	94.38	68.24	77.22	70.11	79.03	27.56	9.79
BiseNet	94.81	67.86	77.6	70.59	79.71	24.87	8.95
DASNet	95.08	69.12	79.42	72.35	81.23	115.79	18.72
MFGAN	95.7	73.92	81.14	75.73	83.15	52.82	11.53
TFI-GR	95.95	71.67	81.06	75.56	83.52	38.79	13.91
SAGNet	95.96	75.24	81.40	76.76	83.99	48.94	19.34
ABMFNet	**96.15**	**76.36**	**81.81**	**77.69**	**84.65**	66.17	22.35

Figure 14 shows the prediction diagrams of various models on the Google Earth images in Double Time. Through the comparison of the renderings, it can be found that although the other 13 deep learning and changing detection algorithms can basically determine the changing areas, there is still a situation of detection misunderstanding or missed inspection. Especially in the coding stage, algorithms of upper samples such as FC_EF, FC-SIAM-Diff, FC-SIAM-Conc, thanks to the new fusion method, result in SNUNet, TCDNet, and MFGAN having achieved good results.

TFI-GR performed well in it, and it can extract and fuse different characteristics of different features. The addition of attention mechanisms has made SAGNet more effective in identifying change areas. However, compared with the predictive diagram of ABMFNET, these methods have not solved the problem of attention mechanism and multi-scale fusion, so there are still some errors in the prediction of the edge details. The ABMFNET structure we use adds three modules: CFM, DEFPN, and AMFFM on the basis of ResNet_34. It pays targeted attention to changing areas and constant areas, and distinguish them. In contrast, our method has more effectively identified the change area, especially the optimization of the edge details, making the prediction results closer to the real label. In summary, ABMFNet effectively addresses some of the issues existing in traditional algorithms in change detection by introducing attention mechanisms and multi-scale fusion, thereby improving the recognition ability of changing areas. It particularly demonstrates significant advantages in handling edge details.

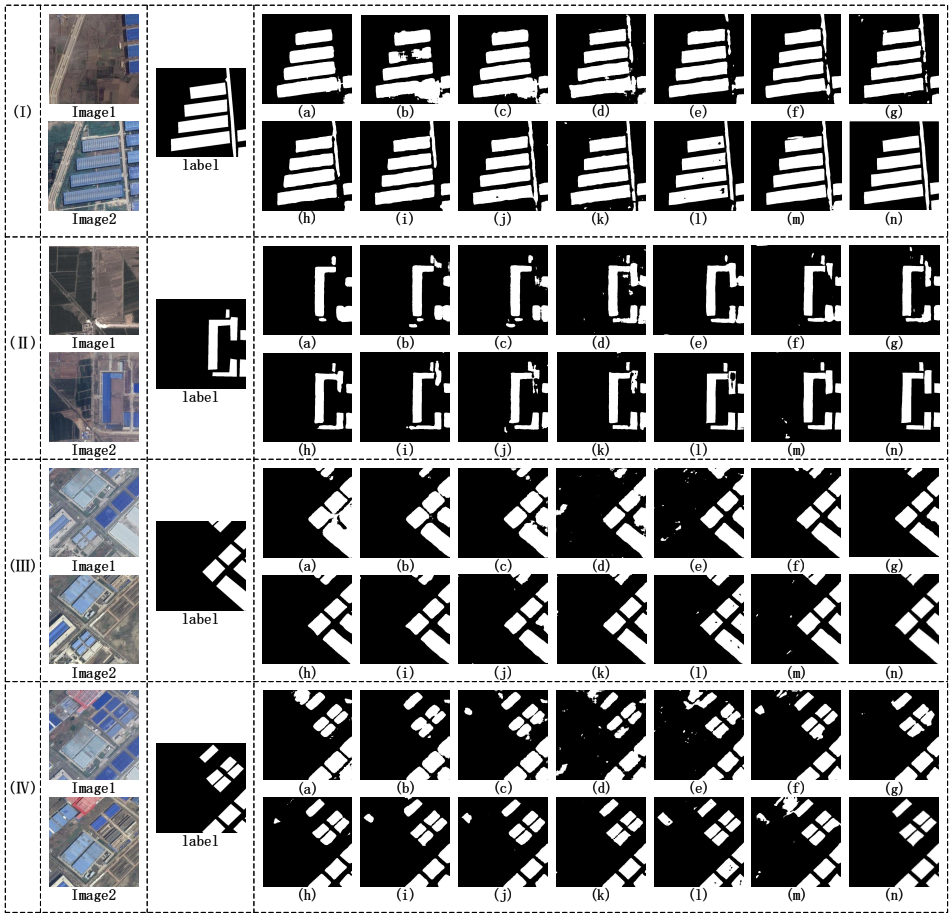

Figure 14. The prediction graphs of several algorithms are compared on the BICD dataset. Different remote sensing images are shown in (**I**–**IV**). Image 1 and Image 2 are bitemporal remote sensing images from different periods. (**a**–**n**) represent the prediction maps of FC_EF, FC-Siam-Diff, FC-Siam-Conc, SNUNet, ChangeNet, HRNet, DeepLabV3+, TCDNet, BiseNet, DASNet, MFGAN, TFI-GR, SAGNet, and ABMFNet, respectively.

3.5.2. Generalization Experiments on LEVIR-CD Dataset and BCDD Dataset

To validate the generalization ability of ABMFNet across different datasets, we conducted comparative experiments using the public datasets LEVIR-CD and BCDD, comparing them with recent change detection algorithms. The experimental results presented in Tables 6 and 7 clearly demonstrate that the performance of deep learning algorithms on the LEVIR-CD and BCDD datasets has been consistently improving with model optimization. In our experiments, the inclusion of multi-scale fusion modules and attention mechanisms led to the most significant improvement in network performance. Specifically, on the LEVIR-CD dataset, our model performed remarkably well, achieving pixel accuracy (PA), recall (RC), precision (PR), F1 score, and mean Intersection over Union (MIoU) of 98.15%, 82.06%, 83.02%, 81.57%, and 85.83%, respectively. Similarly, on the BCDD dataset, our model also achieved satisfactory results, with metrics including PA of 97.38%, RC of 80.36%, PR of 79.67%, F1 of 77.91%, and MIoU of 84.54%. These results fully demonstrate the outstanding performance and generalization ability of the proposed ABMFNet model across different datasets. By incorporating multi-scale fusion modules and attention mechanisms, our model can better capture the dataset's feature information and achieve remarkable performance in change detection tasks. These achievements provide valuable insights and references for future research in remote sensing image change detection.

Table 6. Different algorithm experiments on the LEVIR-CD dataset (bold numbers represent the optimal results).

Method	PA (%)	RC (%)	PR (%)	F1 (%)	MIoU (%)
ChangeNet	97.59	79.57	76.30	76.51	82.47
FC_EF	97.73	77.08	77.43	75.75	82.54
TCDNet	97.66	77.24	79.36	77.26	82.93
BiseNet	97.75	77.75	79.97	77.91	83.38
SegNet	97.91	78.18	79.93	77.52	83.53
MFGAN	97.82	76.85	81.61	78.34	83.59
FC-Siam-Diff	97.99	78.15	80.74	78.60	84.24
FC-Siam-Conc	97.96	81.09	80.73	79.59	84.56
SAGNet	97.98	77.01	84.87	79.62	84.63
HRNet	98.13	78.37	82.17	79.14	84.78
PSPNet	98.03	80.33	81.43	79.87	84.79
FCN8s	98.11	79.29	82.18	79.71	84.96
DeepLabV3+	98.07	80.08	82.13	80.12	85.06
TFI-GR	**98.23**	79.35	81.17	79.4	85.15
SNUNet	98.16	82.44	82.21	81.13	85.58
ABMFNet	98.15	**82.06**	**83.02**	**81.57**	**85.83**

Table 7. Different algorithm experiments on the BCDD dataset (bold numbers represent the optimal results).

Method	PA (%)	RC (%)	PR (%)	F1 (%)	MIoU (%)
FC-Siam-Conc	91.93	76.55	58.06	62.05	72.31
FC-Siam-Diff	94.06	71.99	63.97	64.61	75.41
FC_EF	95.77	65.83	70.04	65.13	77.14
ChangeNet	95.39	73.52	66.77	67.02	77.56
BIT [69]	95.99	71.62	73.08	69.57	78.95
SNUNet	96.01	77.38	69.86	70.45	79.47
DASNet	96.15	72.23	74.36	71.24	80.72
MFGAN	96.95	74.52	75.08	73.01	81.77
TCDNet	96.97	74.83	75.47	73.12	81.79
BiseNet	97.26	75.14	74.61	73.17	82.11
DeepLabV3+	97.31	74.41	75.68	73.31	82.16
SAGNet	97.09	75.96	75.08	73.91	82.50
TFI-GR	97.22	76.37	77.22	74.91	83.04
ABMFNet	**97.38**	**80.36**	**79.67**	**77.82**	**84.54**

Figure 15 illustrates three classic images selected from the LEVIR-CD dataset, showcasing the predictive performance of different algorithms. These images represent various scenes and difficulty levels: (I) depicts a relatively simple background with a significant number of changing buildings in dual-temporal images; (II) presents an image with moderate interference and a suitable amount of changing areas; and (III) shows an image with fewer changing building areas, but strong interference in the background. Observing the predictive results of different algorithms, we find serious issues in small object detection and edge detail accuracy, as well as ambiguous predictions of changing areas. Particularly for type (III) images with stronger interference, the errors of other algorithms are more pronounced. We have highlighted specific misjudged areas in the images using red boxes. In comparison to other algorithms, although SNUNet performs well in terms of metrics, its effectiveness in detecting object details is not satisfactory. With the assistance of attention mechanisms, TFI-GR shows some improvement in detail and edge detection. ABMFNet demonstrates fewer errors and closer approximation to the ground truth labels in predictions. Even when facing significant interference, our model accurately predicts the locations of changing areas and refines edge details.

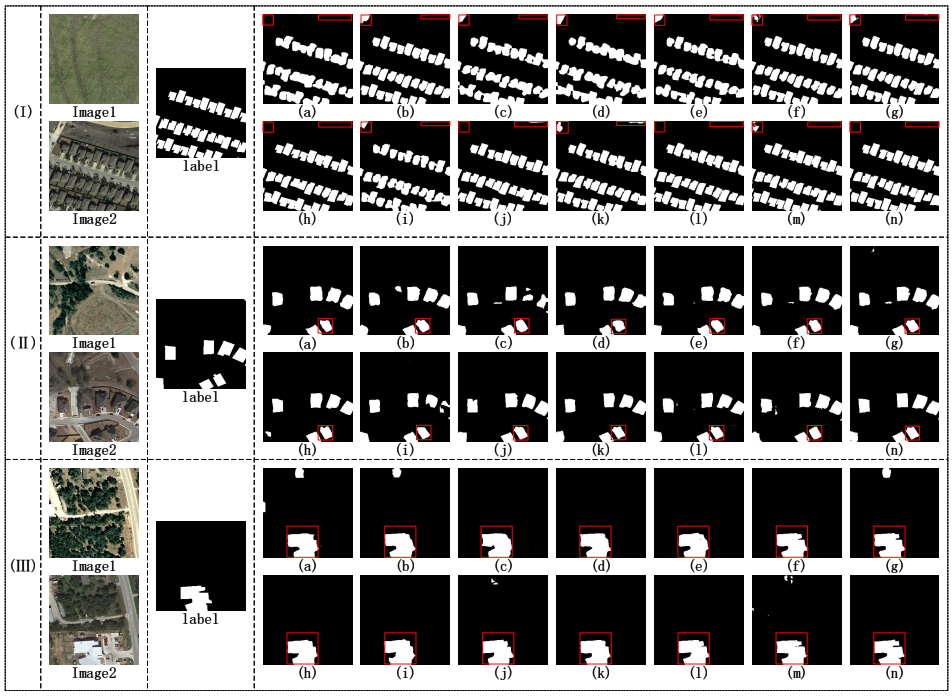

Figure 15. Several deep learning algorithms are compared on the LEVIR-CD dataset. The remote sensing images of different periods are shown in (**I**–**III**). (**a**–**n**) represent the prediction maps of ChangeNet, FC_EF, TCDNet, BiseNet, MFGAN, FC-Siam-Diff, FC-Siam-Conc, SAGNet, DASNet, HRNet, DeepLabV3+, TFI-GR, SNUNet, and ABMFNet, respectively.

As shown in Figure 16 (where red indicates missed detections and blue represents false alarms), it can be observed that, under the interference of similar objects, most models exhibit false detection issues, with FC-Siam-Conc, FC-Siam-Diff, and FC_EF showing relatively severe false detections. Even though SAGNet and TFI-GR incorporate attention mechanisms, the fusion results are not satisfactory, resulting in detection errors as well. our approach pays more attention to changes in regions of interest and effectively handles the interference of pseudo-changes, which is particularly prominent in the BCDD dataset. This dataset only requires the identification of changes in buildings, while other changes such as vehicles and roads can be considered pseudo-changes. It is evident from the prediction images that ABMFNet achieves good change detection results even in the presence of pseudo-changes interference. In conclusion, our model exhibits good performance, robustness, and generalization ability on both the LEVIR-CD and BCDD datasets.

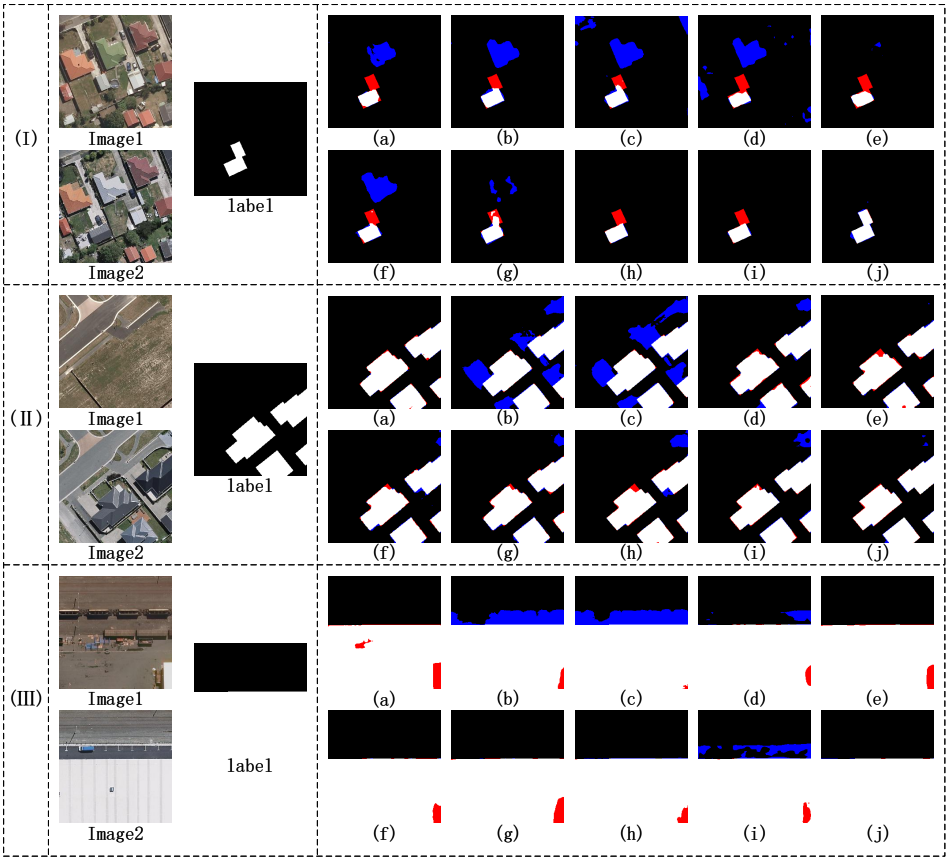

Figure 16. Comparing several deep learning algorithms on the BCDD dataset. The remote sensing images of different periods are shown in (**I**–**III**). (**a**–**j**) represent the prediction maps of FC_EF, FC-Siam-Diff, FC-Siam-Conc, ChangeNet, DASNet, MFGAN, TCDNet, SAGNet, TFI-GR, and ABMFNet (color interpretations are provided in Section 3.5.2).

4. Discussion

For remote sensing change detection tasks, this paper introduces a multi-scale feature fusion Siamese network (ABMFNet) based on an attention mechanism. The network comprises three modules that aid in the training process. To bridge the spatial disparities between low-level and high-level features, we introduce a cross-scale fusion module (CFM) and a bottom-up difference enhancement feature pyramid structure (DEFPN). These modules effectively integrate feature differences at various levels, ensuring the restoration of original image detail features during subsequent upsampling processes. Additionally, we introduce a feature enhancement module (FEM) to strengthen the expression of the feature pyramid, thereby improving inference speed and achieving state-of-the-art performance. What sets our model apart is the attention-based multi-scale feature fusion module (AMFFM). It not only addresses insufficient feature fusion and connection between different-scale feature layers, but also enables the model to automatically learn and select important features or regions in the image, focusing more attention on these areas. On the BICD, LEVIR-CD, and BCDD datasets, ABMFNet demonstrates superior performance in change area recognition and interference resistance compared to other state-of-the-art algorithms. However, this study solely focuses on identifying change regions and targets. In the future, we aim to delve deeper into identifying multi-category change regions. Additionally, in the model comparison section, although ABMFNet achieves promising results, its model complexity poses challenges in reducing model complexity while maintaining detection accuracy and effectiveness. Going forward, we also plan to further explore semi-supervised or unsupervised learning to facilitate the broader application of remote sensing change detection.

5. Conclusions

This paper presents a method for high-resolution remote sensing image change detection tasks based on the multi-scale feature fusion Siamese network (ABMFNet). By introducing key technologies such as the attention-based multi-scale fusion module (AMFFM), cross-scale fusion module (CFM), and bottom-up difference enhancement feature pyramid structure (DEFPN), ABMFNet successfully improves the accuracy and efficiency of change detection. On the BICD, LEVIR-CD, and BCDD datasets, ABMFNet achieves F1 scores of 77.69%, 81.57%, and 77.91%, respectively, as well as MIoU evaluation metrics of 84.65%, 85.84%, and 84.54%, respectively. Experimental results demonstrate the outstanding performance of ABMFNet on these datasets, particularly in the recognition of change areas and resistance to interference, where ABMFNet outperforms other algorithms.

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

Funding: This research was funded by the National Natural Science Foundation of PR China of grant number 42075130.

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

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

References

1. Ren, H.; Xia, M.; Weng, L.; Hu, K.; Lin, H. Dual-Attention-Guided Multiscale Feature Aggregation Network for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 4899–4916. [CrossRef]
2. Chughtai, A.H.; Abbasi, H.; Ismail, R.K. A review on change detection method and accuracy assessment for land use land cover. *Remote Sens. Appl. Soc. Environ.* **2021**, *22*, 100482. [CrossRef]
3. Wang, Z.; Xia, M.; Weng, L.; Hu, K.; Lin, H. Dual Encoder–Decoder Network for Land Cover Segmentation of Remote Sensing Image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2372–2385. [CrossRef]
4. Song, Y.; Jing, Z.; Li, M. Siamese u-net with attention mechanism for building change detection in high-resolution remote sensing images. In Proceedings of the International Conference on Aerospace System Science and Engineering, Shanghai, China, 14–16 July 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 487–503.
5. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A deep multitask learning framework coupling semantic segmentation and fully convolutional lstm networks for urban change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, 59, 7651–7668. [CrossRef]
6. Yang, J.; Weisberg, P.J.; Bristow, N.A. Landsat remote sensing approaches for monitoring long-term tree cover dynamics in semi-arid woodlands: Comparison of vegetation indices and spectral mixture analysis. *Remote Sens. Environ.* **2012**, *119*, 62–71. [CrossRef]
7. Isaienkov, K.; Yushchuk, M.; Khramtsov, V.; Seliverstov, O. Deep learning for regular change detection in ukrainian forest ecosystem with Sentinel-2. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 364–376. [CrossRef]
8. Sublime, J.; Kalinicheva, E. Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the Tohoku tsunami. *Remote Sens.* **2019**, *11*, 1123. [CrossRef]
9. Quarmby, N.A.; Cushnie, J.L. Monitoring urban land cover changes at the urban fringe from spot hrv imagery in south-east England. *Int. J. Remote Sens.* **1989**, *10*, 953–963. [CrossRef]
10. Howarth, P.J.; Wickware, G.M. Procedures for change detection using landsat digital data. *Int. J. Remote Sens.* **1981**, *2*, 277–291. [CrossRef]
11. Ludeke, A.K.; Maggio, R.C.; Reid, L.M. An analysis of anthropogenic deforestation using logistic regression and gis. *J. Environ. Manag.* **1990**, *31*, 247–259. [CrossRef]
12. Nielsen, A.A.; Conradsen, K.; Simpson, J.J. Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sens. Environ.* **1998**, *64*, 1–19. [CrossRef]
13. Nielsen, A.A. The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data. *IEEE Trans. Image Process.* **2007**, *16*, 463–478. [CrossRef]
14. Zhang, H.; Gong, M.; Zhang, P.; Su, L.; Shi, J. Feature-level change detection using deep representation and feature change analysis for multispectral imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1666–1670. [CrossRef]
15. Malila, W.A. Change vector analysis: An approach for detecting forest changes with landsat. LARS Symposia 1980; p. 385. Available online: http://docs.lib.purdue.edu/lars_symp/385 (accessed on 15 July 2023).

16. Kuncheva, L.I.; Faithfull, W.J. Pca feature extraction for change detection in multidimensional unlabeled data. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *25*, 69–80. [CrossRef]
17. 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]
18. Ma, L.; Li, M.; Blaschke, T.; Ma, X.; Tiede, D.; Cheng, L.; Chen, Z.; Chen, D. Object-based change detection in urban areas: The effects of segmentation strategy, scale, and feature space on unsupervised methods. *Remote Sens.* **2016**, *8*, 761. [CrossRef]
19. Zhang, Y.; Peng, D.; Huang, X. Object-based change detection for vhr images based on multiscale uncertainty analysis. *IEEE Geosci. Remote Sens. Lett.* **2017**, *15*, 13–17. [CrossRef]
20. Zhang, C.; Li, G.; Cui, W. High-resolution remote sensing image change detection by statistical-object-based method. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2440–2447. [CrossRef]
21. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
22. Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; Qin, B. Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1, Long Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 1555–1565.
23. Yoon, K.; Yacine, J.; David, S.; Rush, A.M. Character-aware neural language models. In Proceedings of the 13th AAAI Conference on Artificial Intelligence, Phoenix, AR, USA, 12–13 February 2016.
24. Lei, T.; Zhang, Q.; Xue, D.; Chen, T.; Meng, H.; Nandi, A.K. End-to-end change detection using a symmetric fully convolutional network for landslide mapping. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3027–3031.
25. Li, X.; Yuan, Z.; Wang, Q. Unsupervised deep noise modeling for hyperspectral image change detection. *Remote Sens.* **2019**, *11*, 258. [CrossRef]
26. Xu, Q.; Chen, K.; Zhou, G.; Sun, X. Change capsule network for optical remote sensing image change detection. *Remote Sens.* **2021**, *13*, 2646. [CrossRef]
27. Chen, K.; Dai, X.; Xia, M.; Weng, L.; Hu, K.; Lin, H. MSFANet: Multi-Scale Strip Feature Attention Network for Cloud and Cloud Shadow Segmentation. *Remote Sens.* **2023**, *15*, 4853. [CrossRef]
28. Ji, H.; Xia, M.; Zhang, D.; Lin, H. Multi-Supervised Feature Fusion Attention Network for Clouds and Shadows Detection. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 247. [CrossRef]
29. Dai, X.; Chen, K.; Xia, M.; Weng, L.; Lin, H. LPMSNet: Location Pooling Multi-Scale Network for Cloud and Cloud Shadow Segmentation. *Remote Sens.* **2023**, *15*, 4005. [CrossRef]
30. Lecun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.
31. Ren, W.; Wang, Z.; Xia, M.; Lin, H. MFINet: Multi-Scale Feature Interaction Network for Change Detection of High-Resolution Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1269. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
33. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Gong, M.; Zhao, J.; Liu, J.; Miao, Q.; Jiao, L. Change detection in synthetic aperture radar images based on deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 125–138. [CrossRef]
36. Zhan, Y.; Fu, K.; Yan, M.; Sun, X.; Wang, H.; Qiu, X. Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1845–1849. [CrossRef]
37. Wang, D.; Chen, X.; Jiang, M.; Du, S.; Xu, B.; Wang, J. Ads-net: An attention-based deeply supervised network for remote sensing image change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *101*, 102348.
38. Yin, H.; Weng, L.; Li, Y.; Xia, M.; Hu, K.; Lin, H.; Qian, M. Attention-guided siamese networks for change detection in high resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103206. [CrossRef]
39. Li, L.; Lv, M.; Jia, Z.; Ma, H. Sparse representation-based multi-focus image fusion method via local energy in shearlet domain. *Sensors* **2023**, *23*, 2888. [CrossRef]
40. Zhang, X.; Li, W.; Gao, C.; Yang, Y.; Chang, K. Hyperspectral pathology image classification using dimension-driven multi-path attention residual network. *Expert Syst. Appl.* **2023**, *230*, 120615. [CrossRef]
41. Zhang, X.; Li, Q.; Li, W.; Guo, Y.; Zhang, J.; Guo, C.; Chang, K.; Lovell, N.H. FD-Net: Feature Distillation Network for Oral Squamous Cell Carcinoma Lymph Node Segmentation in Hyperspectral Imagery. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1552–1563. [CrossRef]
42. Ma, C.; Weng, L.; Xia, M.; Lin, H.; Qian, M.; Zhang, Y. Dual-branch network for change detection of remote sensing image. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106324. [CrossRef]
43. Chu, S.; Li, P.; Xia, M. Mfgan: Multi feature guided aggregation network for remote sensing image. *Neural Comput. Appl.* **2022**, *34*, 10157–10173. [CrossRef]

44. Ding, L.; Xia, M.; Lin, H.; Hu, K. Multi-Level Attention Interactive Network for Cloud and Snow Detection Segmentation. *Remote Sens.* **2024**, *16*, 112. [CrossRef]
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
46. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
47. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
48. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
49. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
50. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 431–435. [CrossRef]
51. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 30 June 2016; pp. 2818–2826.
52. Purkait, P.; Zhao, C.; Zach, C. SPP-Net: Deep absolute pose regression with synthetic views. *arXiv* **2017**, arXiv:1712.03452.
53. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
54. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
55. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
56. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
57. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
58. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
59. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
60. Daudt, R.C.; Saux, B.L.; 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.
61. 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.* **2021**, *19*, 1–5. [CrossRef]
62. Varghese, A.; Gubbi, J.; Ramaswamy, A.; Balamuralidhar, P. ChangeNet: A Deep Learning Architecture for Visual Change Detection. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
63. Qian, J.; Xia, M.; Zhang, Y.; Liu, J.; Xu, Y. Tcdnet: Trilateral change detection network for google earth image. *Remote Sens.* **2020**, *12*, 2669. [CrossRef]
64. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1194–1206. [CrossRef]
65. Li, Z.; Tang, C.; Wang, L.; Zomaya, A.Y. Remote sensing change detection via temporal feature interaction and guided refinement. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
66. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
67. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818;
68. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
69. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]