remote sensing

MDPI

Article

AMFNet: Attention-Guided Multi-Scale Fusion Network for Bi-Temporal Change Detection in Remote Sensing Images

Zisen Zhan [1], Hongjin Ren [1], Min Xia [1,*], Haifeng Lin [2], Xiaoya Wang [1,3] and Xin Li [2]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202183250030@nuist.edu.cn (Z.Z.); 202212220006@nuist.edu.cn (H.R.); ws804641@student.reading.ac.uk (X.W.)
[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210000, China; haifeng.lin@njfu.edu.cn (H.L.); csxinli@njfu.edu.cn (X.L.)
[3] Department of Computer Science, University of Reading, Whiteknights, Reading RG66DH, UK
* Correspondence: xiamin@nuist.edu.cn

Abstract: Change detection is crucial for evaluating land use, land cover changes, and sustainable development, constituting a significant component of Earth observation tasks. The difficulty in extracting features from high-resolution images, coupled with the complexity of image content, poses challenges for traditional change detection algorithms in terms of accuracy and applicability. The recent emergence of deep learning methods has led to substantial progress in the field of change detection. However, existing frameworks often involve the simplistic integration of bi-temporal features in specific areas, lacking the fusion of temporal information and semantic details in the images. In this paper, we propose an attention-guided multi-scale fusion network (AMFNet), which effectively integrates bi-temporal image features and diverse semantics at both the encoding and decoding stages. AMFNet utilizes a unique attention-guided mechanism to dynamically adjust feature fusion, enhancing adaptability and accuracy in change detection tasks. Our method intelligently incorporates temporal information into the deep learning model, considering the temporal dependency inherent in these tasks. We decode based on an interactive feature map, which improves the model's understanding of evolving patterns over time. Additionally, we introduce multi-level supervised training to facilitate the learning of fused features across multiple scales. In comparison with different algorithms, our proposed method achieves F1 values of 0.9079, 0.8225, and 0.8809 in the LEVIR-CD, GZ-CD, and SYSU-CD datasets, respectively. Our model outperforms the SOTA model, SAGNet, by 0.69% in terms of F1 and 1.15% in terms of IoU on the LEVIR-CD dataset, by 2.8% in terms of F1 and 1.79% in terms of IoU on the GZ-CD dataset, and by 0.54% in terms of F1 and 0.38% in terms of IoU on the SYSU-CD dataset. The method proposed in this study can be applied to various complex scenarios, establishing a change detection method with strong model generalization capabilities.

Keywords: change detection; remote sensing image; deep learning; multi-scale supervised

Citation: Zhan, Z.; Ren, H.; Xia, M.; Lin, H.; Wang, X.; Li, X. AMFNet: Attention-Guided Multi-Scale Fusion Network for Bi-Temporal Change Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1765. https://doi.org/10.3390/rs16101765

Academic Editors: Sananda Kundu and Arun Mondal

Received: 5 March 2024
Revised: 5 May 2024
Accepted: 13 May 2024
Published: 16 May 2024

1. Introduction

Land surface change monitoring is an important application of remote sensing image change detection [1], extensively utilized across various domains, including land management [2], disaster evaluation [3], agricultural analysis [4], landslide detection [5], and ecosystem surveillance [6]. However, the increasing pace of urbanization and shifts in ecological environments have significantly increased the complexity of land surfaces. This increase in complexity is marked by a diversification of background attributes and a concentration of dynamic features, such as buildings and vegetation, thereby amplifying the challenges associated with change detection tasks [7,8]. In this process, each pixel is assigned a binary label, with 0 indicating no change and 1 indicating change. Features extracted from photos taken at the same location but at different periods are used to define

these categories [4]. Research in the domain of change detection commenced with a notable study by West et al. [9] in 1977, employing differential techniques in remote sensing imagery to detect changes, signaling the inception of this field of study. Pixel-based techniques, including image algebra, picture manipulation, and classification comparison, have been used in this field traditionally. The algebraic method involves generating a different image through the calculation of pixel information in bi-temporal remote sensing images. Subsequently, an appropriate threshold is applied to determine changed and unchanged areas [10], employing methods like the ratio method [11], regression method [12], change vector analysis [13], etc. In 2020, Du et al. [14] introduced a method known as logic-verified change vector analysis (LV-CVA), an advancement of change vector analysis (CVA). This method integrates logical reasoning and decision-making processes and incorporates additional temporal imagery to accurately identify errors in CVA. The core concept of image transformation methods lies in the reselection and recombination of spectral bands or the application of projection transformations. By extracting the fundamental features of the data and using these more representative features, this approach seeks to improve detection efficiency and precision. Notable techniques in this category include principal component analysis (PCA) [15,16] and multivariate change detection (MCD) [17]. Unsupervised change detection is commonly achieved by algebra-based and transformation-based techniques [18,19]. These approaches rely on model assumptions or comparison rules to identify change areas; however, they are not without limitations. To improve detection performance, many researchers have turned unsupervised methods into supervised methods. Among them, the classification comparison method is a popular supervised remote sensing change detection method [20]. The photos from two different periods are classified independently, and the position and kind of changes are identified by comparing the classification results pixel by pixel [21]. Arithmetic and transformation-based methods are less effective in high-resolution images since they mostly rely on empirical design. Juan et al. [22] developed a detection algorithm that includes both pixel and object representations. After segmenting large images into overlapping sub-images, supervised sub-image retrieval is employed to identify changed regions containing artificial objects. Wan et al. [23] aimed to reduce false positives from region shifts and object misplacements in traditional methods using synthetic aperture radar (SAR) and optical images as simultaneous data sources. With this technique, multi-temporal data from several sensors may be used effectively and interference-free. In high-resolution data processing using object-oriented methods, the focus has shifted from individual pixels to segmented objects. Objects, unlike pixels, provide richer contextual information [24]. In terms of accuracy, object-oriented change detection techniques, such as post-classification [25] and multiple and multi-scale classification [26], typically perform better than pixel-level techniques. However, these approaches can lose detail and vary in accuracy due to their reliance on differential images and superpixel generation. To address this, Han et al. [27] applied a weighted Dempster–Shafer theory (wDST) fusion approach, integrating various pixel-based detections to yield more stable object-based change detection outcomes.

The advent of the big data age and improvements in high-performance computers are the main causes of the recent explosion in deep learning [28–31]. Its ability to effectively extract deep features has led to its application in remote sensing image analysis [32–34]. Convolutional neural networks (CNNs), a fundamental deep learning framework, are adept at autonomously learning the complex spatial–spectral characteristics of remote sensing imagery [35–37]. However, due to their high computational needs and stringent input–output size restrictions, the fully connected layers in conventional CNN change detection approaches encounter difficulties. To address this, fully convolutional networks (FCNs) have been introduced. FCNs employ end-layer convolution operations to shift from image-level to pixel-level classification [38], making them compatible with pixel-based change detection. Furthermore, FCN-based techniques have advanced remote sensing segmentation, leading to diverse neural network models and significant breakthroughs. Notably, Li et al. [39] enhanced FCNs with multi-scale convolution modules, demonstrating

the efficacy of multi-scale features in refining change detection in high-resolution imagery. U-Net [40] enhances the FCN framework by integrating an innovative feature fusion technique. It merges features at the channel level, creating more impactful characteristics than those generated by FCN. The transformer technique from natural language processing (NLP) was applied to the field of computer vision (CV) by Chen et al. [41]. In order to achieve comprehensive end-to-end change detection, Xing et al. [42] combined a pyramid-shaped decoder. This innovative approach has resulted in a significantly smaller model size, making it more suitable for peripheral devices and industrial uses. Chen et al. [43] introduced an inventive use of a self-attention mechanism for change detection, specifically designed to model spatio-temporal relationships, with an emphasis on the collection of neighboring scale information.

Currently, neural networks encounter several obstacles in the field of change detection [44–47]. (1) Common techniques frequently combine bi-temporal remote sensing pictures based merely on the channel aspect for extracting features. While deep neural networks have the capability to implicitly discern differences between images, this method frequently falls short of effectively capturing essential differences [48]. (2) Most change detection networks prioritize identifying key features within each individual time phase during bi-temporal feature fusion, thus overlooking critical elements of interaction across bi-temporal dimensions. Improving the extraction of information from hyper-spectral image inputs [49], as well as enhancing the integration of semantic information across the same-level feature maps of bi-temporal hyperspectral images, represents a significant area for potential advancements [50,51]. Consequently, an attention-guided multi-scale fusion network is introduced, which is specially designed to integrate feature information across multiple scales and dimensions in an integrated manner. Using the attention mechanism, AMFNet can improve the effectiveness of the feature fusion process and ensure that the information flow between different scales is more accurate and targeted. In addition, by encoding on the interactive feature map, AMFNet further makes full use of bi-temporal information, rather than a single temporal graph, to extract features alone, which enhances the sensitivity and analytical ability of the model to time changes. In the feature encoding stage, we use the weight-sharing systematic down-sampling method to construct a pyramid-shaped multi-scale [52] feature map. This method can effectively extract various features of bi-temporal hyper-spectral remote sensing images simultaneously, including textural, spectral, and spatial features. In this way, the model can capture subtle changes in images at different scales. For two corresponding images at an identical layer, the bi-temporal fusion attention module (BFAM) plays a pivotal role. This module discerns the most significant components among the bi-temporal features, enabling effective integration and localization of the features associated with bi-temporal change objects. We further process the feature maps, which have undergone initial information exchange through BFAM, by applying addition and concatenation. This step is crucial for extracting both differential and comprehensive global information. The bilateral fusion module (BFM) efficiently merges this information. The integrated attention module (IAM) refines this process, focusing on extracting channel and spatial variations within the bi-temporally fused feature maps. The following summarizes the primary contributions of our study:

1. We propose an attention-guided multi-scale fusion network (AMFNet) for change detection in high-resolution remote sensing images. The network makes full use of the abundant features of remote sensing images and optimizes feature interaction and semantic information fusion through an attention mechanism, effectively addressing issues of uncertain target edges and omissions.
2. We propose the bi-temporal fusion attention module (BFAM) and bilateral fusion module (BFM). BFAM can combine channel and spatial attention mechanisms and utilizing temporal information. BFM extracts the differential and global information of bi-temporal features, better pinpointing detailed features and texture characteristics, achieving complementary of information between the two branches.

3. The integrated attention module (IAM) is introduced to allow the network to identify diverse features across spatial and channel dimensions while eliminating and reducing redundant features. It extracts the changing regions as positions with high feature weights, thereby enhancing the network's detection accuracy.
4. Our AMFNet, as shown by comprehensive testing on two datasets for remote sensing image change detection, achieves both robustness and superior accuracy, outperforming other deep learning change detection methods.

2. Materials and Methods

2.1. Proposed Approach

2.1.1. Network Structure

In this study, we propose an attention-guided multiscale fusion network with the goal of improving the accuracy of differentiating between changed and unchanged areas. This network structure consists of one decoding branch and two encoding branches. Initially, each bi-temporal picture is input separately into a Siamese ResNet34 [53] with shared weights, generating multi-level features. Subsequently, for the focused and effective fusion of bi-temporal features, the feature maps at the same hierarchical level in the encoding process are passed through the BFAM to generate corresponding interactive feature maps. Each interactive feature map contains valid semantic information from both temporal instances. In the encoding part, after merging the interacted feature maps on an equal footing, we input the interactive feature maps of corresponding layers into both addition and concatenation operations during the decoding stage. This process extracts diverse semantic information. Both sets of semantic information are then fused through the bilateral fusion module (BFM) and further refined by the integrated attention module (IAM) to enhance feature representation. These two modules, when used consecutively, can effectively achieve the coherence of special extraction and enhancement. In the decoding phase of our network, the feature maps generated from shallow layers are upsampled and then combined with the corresponding interactive feature maps from deeper layers. This process is repeated, layer by layer, until the feature map size matches the original input size. During the decoding process, we adopt a multiscale supervised learning approach to proportionally fuse the cross-entropy losses. The final result is produced through a sigmoid classifier. The overall schematic diagram is illustrated in Figure 1.

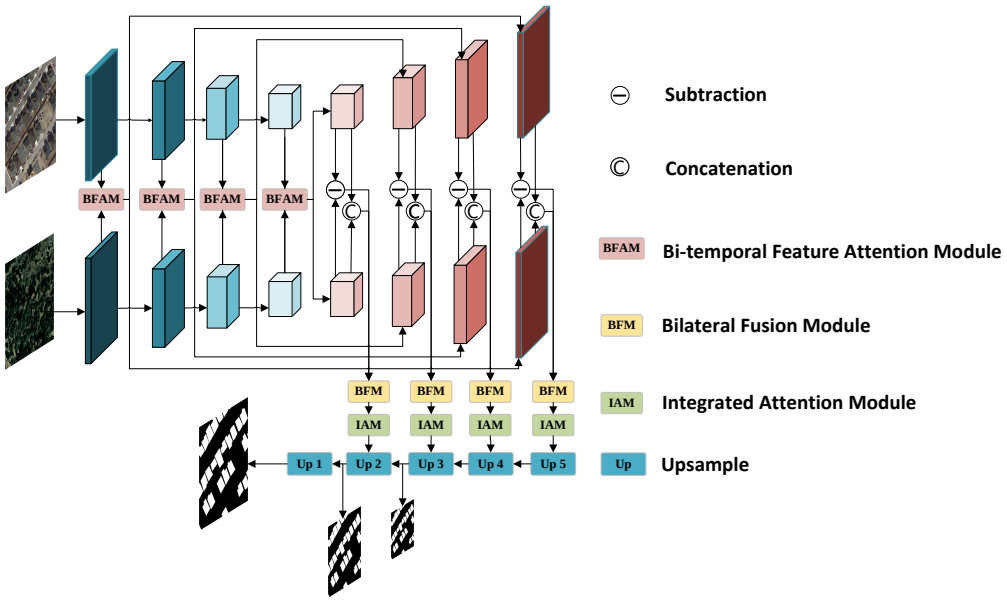

Figure 1. General overview of AMFNet.

2.1.2. Bi-Temporal Feature Attention Module

In the change detection model, bi-temporal feature fusion involves operator fusion methods, convolution fusion methods, and attention fusion methods. Simple operator fusion methods directly add, subtract, or concatenate bi-temporal features for fusion [54–56]. However, noise in bi-temporal features might easily interfere with obtaining reliable detection results using these methods. Convolutional fusion methods enhance bi-temporal features in terms of multi-scale and semantic hierarchies through various convolution operations, reducing the noise interference on bi-temporal features [57]. Typically, the attention fusion approaches use an attention mechanism to effectively fuse bi-temporal characteristics in the channel and spatial dimensions. Convolutional methods concentrate on enhancing bi-temporal characteristics before fusion, while attention refinement approaches focus on improving these characteristics after operator fusion. The temporal information of bi-temporal characteristics is disregarded by both.

As illustrated in Figure 2, we propose a bi-temporal feature attention model (BFAM) to address the problems mentioned previously. BFAM effectively fuses features by utilizing temporal information. The significant portions of the characteristics are identified using channel and spatial attention, and the key sections between the bi-temporal features are identified using temporal information. The channel branch achieves the aggregate of channel information by transmitting the input bi-temporal characteristics across channels via global pooling. The aggregation process can be represented as follows:

$$fc = Concat(Avg(T1\_input), Max(T1\_input), Avg(T2\_input), Max(T2\_input)), \quad (1)$$

where f_c represents the fused channel features. $Concat$ denotes concatenation in the channel dimension. Global average pooling is represented by $Avg(.)$, and global maximum pooling by $Max(.)$. $T1\_input$ and $T2\_input$ are the bi-temporal feature inputs. The fused channel features are then passed through one-dimensional convolutions separately with adaptively determined kernel numbers (k), effectively capturing dependencies between channels. This is followed by the generation of attention weights W_{c1} and W_{c2} for the channels using a nonlinear activation function, softmax. The formulation of the two channel weights can be expressed as follows:

$$W_{c1}, W_{c2} = \frac{e^{Conv_1(f_c)}}{e^{Conv_1(f_c)} + e^{Conv_2(f_c)}}, \frac{e^{Conv_2(f_c)}}{e^{Conv_1(f_c)} + e^{Conv_2(f_c)}}, \quad (2)$$

where $Conv1(.)$ and $Conv2(.)$ represent one-dimensional convolution. The methodology for determining the weights in the spatial dimension parallels that of the channel dimension. Global pooling is applied in the spatial dimension, and weights are obtained using two-dimensional convolution and a nonlinear activation function. In the spatial branch, we identify specific regions between bi-temporal characteristics by calculating the bi-temporal spatial weights W_{S1} and W_{S2}. By integrating bi-temporal channel weights and bi-temporal spatial weights, comprehensive bi-temporal weights are obtained, highlighting the essential parts among bi-temporal features. Subsequently, these bi-temporal weights are multiplied with the bi-temporal features for an effective fusion of these features. The resulting output is processed through a 1×1 convolution and a depth-wise separable convolution, characterized by a lower parameter count, and is then summed with the original input. The output can be formalized as follows:

$$\begin{aligned} f_{T1\_output} &= f^{ds}(f^{1\times1}(W_{C1} + W_{S1}) \otimes f_{T1\_input})) + f_{T1\_input}, \\ f_{T2\_output} &= f^{ds}(f^{1\times1}(W_{C2} + W_{S2}) \otimes f_{T2\_input})) + f_{T2\_input}, \end{aligned} \quad (3)$$

where $f^{1\times1}$ represents a set of 1×1 convolutions, batch normalization, and ReLU activation functions, and f^{ds} stands for depth-wise separable convolution [58], which encompasses depth-wise convolution and point-wise convolution. $\otimes$ stands for the multiplication

operation. This approach significantly reduces the number of parameters while ensuring a larger receptive field.

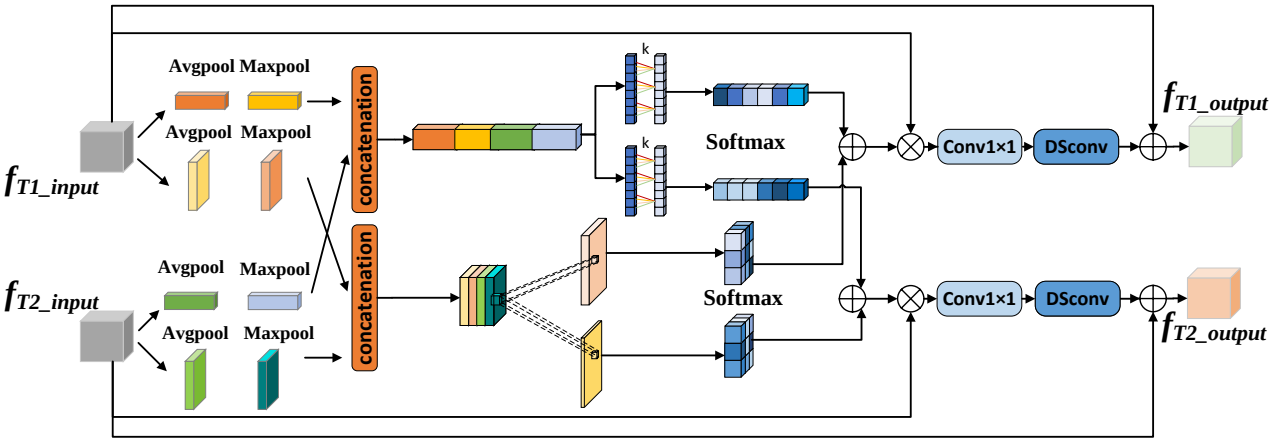

Figure 2. Bi-temporal feature attention module.

2.1.3. Bilateral Fusion Module

In change detection, the commonly used methods for extracting change information involve addition, subtraction, and concatenation operations [47]. However, these simple operator operations for extracting change features often lead to redundant information, significantly reducing the robustness of change detection. Moreover, deep-level semantic information in the network is abstract, and direct simple operations can result in misjudgments and omissions [59]. To reasonably utilize the difference information and global information generated from the subtraction and concatenation of bi-temporal feature maps, we propose a BFM module, as shown in Figure 3. This module aims to precisely locate changed positions from the abstract bi-temporal feature information in the deep network, further revealing details and texture features of the changes.

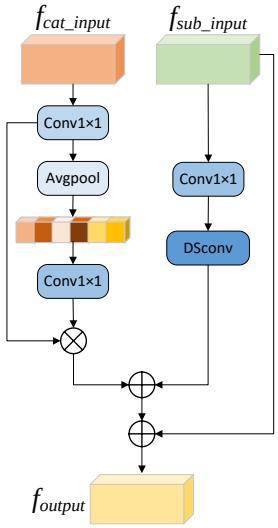

Figure 3. Bilateral fusion module.

Two branches make up the module: the right branch for extracting difference information and the left branch for extracting global information. Since the initial difference information is obtained by subtracting bi-temporal feature maps, the size of f_sub is $C \times H \times W$. It is then processed through a 1×1 convolution and a depth-wise separable convolution [58] to further extract difference features. For the left branch, the initial global

information is obtained by concatenating bi-temporal feature maps along the channel dimension, resulting in f_cat with a size of $2C \times H \times W$. It first passes through a 1×1 convolution to obtain an output of size $C \times H \times W$ and then undergoes global average pooling to obtain a feature map of size $C \times 1 \times 1$. After further extracting features through a 1×1 convolution from the weight feature map, it is multiplied by the feature map of size $C \times H \times W$ before global pooling to obtain more accurate global information. The ultimate output, f_out, is obtained by fusing the characteristics of the two branches and adding the result to f_sub.

Through a sequence of convolution and pooling procedures, the module better extracts change features. The use of a residual structure guides the difference between information and global information, which mutually influence each other, resulting in finer granularity in extracting the edges of change regions. The following are the mathematical expressions:

$$f\_out = f^{1\times1}(Avgpool(f^{1\times1}(f\_cat)))^{1\times1}(f\_sub) + f\_sub. \quad (4)$$

2.1.4. Integrated Attention Module

Remote sensing images, especially high-resolution ones, frequently encompass extensive data volumes and intricate content details. However, our focus is on the change information between bi-temporal images, primarily including buildings, vegetation, rivers, and infrastructure [55]. The characteristics of the changed targets are often not obvious [40], and the complex background interferes with the target areas [25]. For each pixel label, it is challenging for the network to classify and recognize correctly, which may lead to misjudgment and omission of target features. Considering that unidirectional axial attention may result in the omission of feature information in other dimensions, we propose a novel attention mechanism module, IAM, which consists of two branches. The first branch guides the initial features, while the second branch integrates spatial attention and channel attention mechanisms, enabling simultaneous focus on channel and spatial change information. The two branches employ skip connections for feature fusion, facilitating mutual guidance between the branches. This adaptive attention mechanism allows for more attention to be placed on the changing areas, highlighting their importance while suppressing the non-changing regions. The output is more refined attention features, with a greater focus on the details of the edges in the changing areas. The structure of this module is displayed in Figure 4.

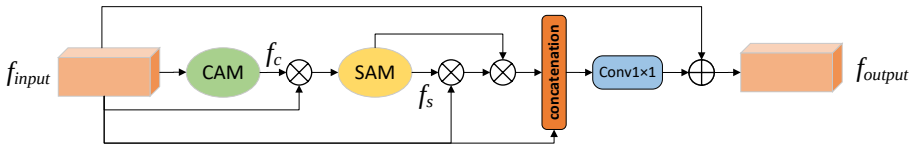

Figure 4. Integrated attention module.

Suppose that input $f_{input} \in R^{C\times H\times W}$, where H and W denote the feature map's height and width, respectively, while C signifies the quantity of channels involved. Initially, the input features are subjected to channel dimension weight resetting by CAM, which involves spatial axis operations on features under each channel. The structure of CAM is illustrated in Figure 5. Specifically, this means condensing the feature map into two vectors of size $C \times 1 \times 1$ by executing both spatial plane max pooling and global average pooling. These two vectors are then input separately into a multi-layer perceptron (MLP) in order to perform classification. The channel attention weights $W_d(d = 1, \ldots, D)$ (where D is the feature dimension and d is the dth feature) are calculated by integrating the MLP outputs and passing them through a sigmoid activation function. These weights also have the size of $C \times 1 \times 1$. Each element of this vector represents the receptive field of the corresponding channel space. The weighted fusion of the channel attention weights W_d and the input feature map f_in results in the process of refining input features in the channel. This

reduces unnecessary channels and highlights those that are important for change detection. The computational formula for the above process can be described as follows:

$$W_d = \sigma(MLP(AvgPool(f_{input})) + MLP(MaxPool(f_{input})), \quad (5)$$

$$f_c = f_{in} \otimes W_d, \quad (6)$$

where σ represents sigmoid activation function. MLP represents multi-layer perceptron operation. $Avg(.)$ and $Max(.)$ represent global average pooling and global maximum pooling, respectively.

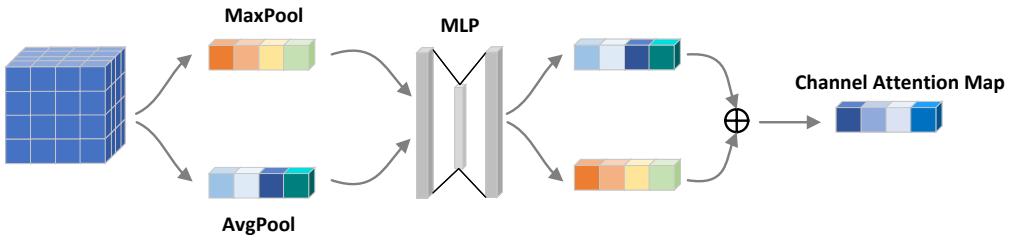

Figure 5. Channel attention module.

SAM begins by applying AvgPool and MaxPool operations along the channel axis to the feature map f_{in}. The structure of SAM is shown in Figure 6. This compression of channel information produces two tensors of size $1 \times H \times W$. These tensors are then concatenated and processed through two consecutive 3×3 convolution operations, which flexibly construct spatial relationships of features. This convolutional operation is equivalent to the receptive field of a 3×3 convolution but with a significantly reduced parameter count. Finally, a sigmoid activation function is introduced to enhance its nonlinear expression, resulting in the spatial attention weight map $W_p(p = 1, \ldots, N)$, where N represents the total number of pixels in each feature map and p denotes the value of the pth pixel in the attention map. Higher and lower weights are, respectively, allocated to pixels that have changed and those that have not changed. The input feature map f_{in} is multiplied element-wise by each element in W_p. Pixels within changing regions are multiplied by higher weights, while pixels within unchanged areas are suppressed by lower weights, achieving spatial refinement. This enables the network to rapidly detect areas undergoing change. The computation formula is as follows:

$$W_p = \sigma(f^{3\times3}(f^{3\times3}([Avgpool(f_c); MaxPool(f_c)]))), \quad (7)$$

$$f_s = f_{in} \otimes W_p, \quad (8)$$

where $f^{3\times3}(\cdot)$ symbolizes a two-dimensional convolution, inclusive of batch normalization and a ReLU activation function, characterized by a 3-unit convolution kernel. $[;]$ denotes the operation of concatenation along the channel dimension. Meanwhile, f_s is indicative of a feature map that has been enhanced through the use of SAM. f_s and the original input undergo a series of multiplication and fusion operations, enhancing feature selection and preserving important information. The concatenated feature map is then passed through a convolutional layer with a kernel size of $f^{1\times1}$. Lastly, the final output feature map, f_output, is obtained by adding the $Conv1 \times 1$ output element-by-element to the original feature map. The formula for the process is specified as follows:

$$S_1 = (f_s \otimes f_{input}) \otimes f_s, \quad (9)$$

$$S_2 = S_1 \; Concat \; f_{input}, \quad (10)$$

$$f_{output} = f^{1\times1}(S_2) \oplus f_{input}. \quad (11)$$

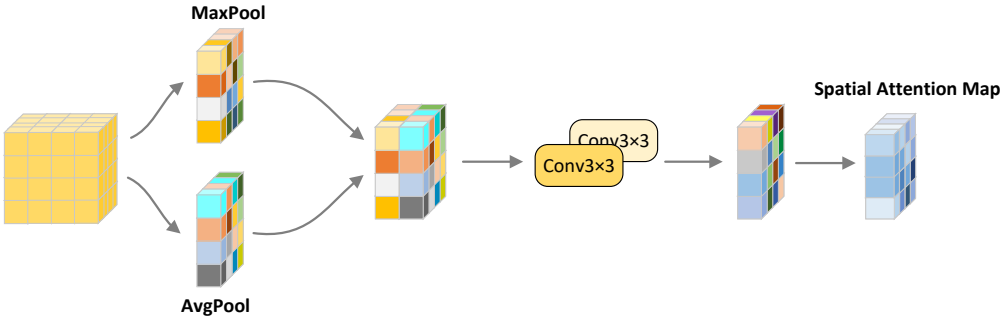

Figure 6. Spatial attention module.

2.2. Datasets

In order to thoroughly confirm the validity of our AMFNet model, we assessed its capability using three distinct remote sensing image change detection datasets: SYSU-CD [60], LEVIR-CD [43], and GZ-CD [61].

2.2.1. SYSU-CD

Sun Yat-sen University created and released the SYSU-CD dataset [60] in 2022, which is a collection of datasets created especially for change detection applications. This dataset is primarily focused on detecting changes in various natural elements. The dataset comprises distinct sets for training, validating, and testing purposes. It is methodically divided in a ratio of 6:2:2, with 12,000 images allocated for training, 4000 for validation, and another 4000 for testing, ensuring comprehensive coverage for machine learning applications. Each image within the dataset maintains a standard resolution of 256 × 256 pixels, suitable for detailed analysis. Figure 7 displays some example.

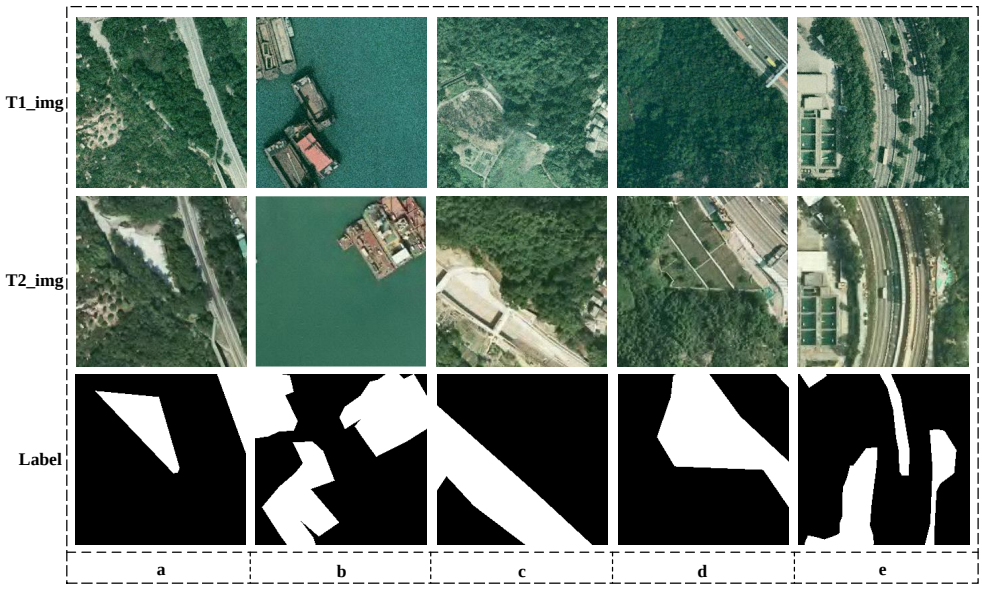

Figure 7. Diagram of the SYSU-CD dataset. (**a**–**e**) correspond to example images in the dataset.

2.2.2. LEVIR-CD

Including very high-quality Google Earth pictures, the LEVIR-CD [43] dataset is a large-scale change detection dataset. The dataset spans a timeframe of 5 to 14 years, with the images captured between 2002 and 2018 primarily focusing on significant transformations in building structures. It includes bi-temporal remote sensing images from 20 distinct areas across multiple cities in Texas, encompassing a wide array of buildings like villas, skyscrapers, small garages, and large warehouses. This diverse collection takes into

consideration changes due to seasonal shifts and varying light conditions. For practical applications, the images are resized to 256 × 256 pixels. Using a 7:1:2 ratio, the dataset is logically split into training, validation, and test sets containing 7120, 1024, and 2048 picture pairings, respectively. The data structure is depicted in Figure 8.

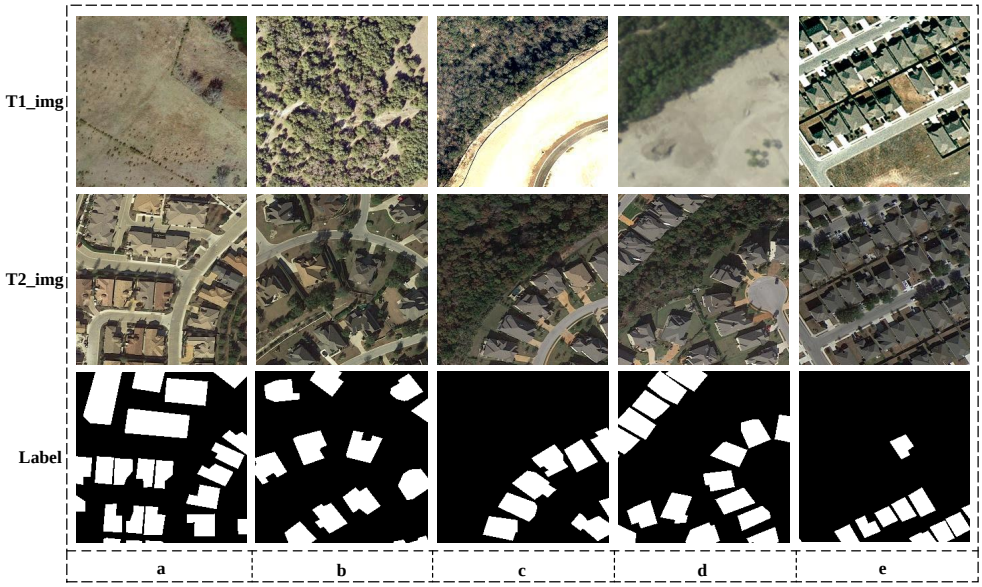

Figure 8. Diagram of the LEVIR-CD dataset. (**a**–**e**) correspond to example images in the dataset.

2.2.3. GZ-CD

The GZ-CD dataset [61], obtained from a suburban area of Guangzhou, China, captures the rapid urban development experienced during its collection, including various alterations in the architecture's shape and size. It showcases a rich diversity of building types, ranging from expansive industrial facilities to compact mobile homes. Given that the bi-temporal images in this dataset maintain consistent resolution, they were specifically utilized for resolution-independent difference experiments within this methodology to evaluate the model's adaptability and generalization skills. Figure 9 displays some examples.

Figure 9. Diagram of the GZ-CD dataset. (**a**–**e**) correspond to example images in the dataset.

2.3. Implementation Details

2.3.1. Evaluation Metrics

To evaluate the performance of our AMFNet model in change detection tasks, we utilized a comprehensive set of evaluation metrics, including precision (PR), recall (RC), overall accuracy (OA), Kappa coefficient ($KAPPA$), intersection over union (IoU), and F1 score ($F1$). These metrics collectively offer a detailed assessment of the model. The primary indicators used for evaluation are the F1 score, which measures the accuracy of the changed category, and the IoU, assessing the overlap between predicted and actual changed areas. Additionally, PR, RC, OA, and $KAPPA$ serve as supporting indicators, enriching the evaluation by providing insights into various aspects of the model's accuracy and reliability. The calculation method of each index is as follows:

$$PR = \frac{TP}{TP + FP}, \tag{12}$$

$$RC = \frac{TP}{TP + FN}, \tag{13}$$

$$F1 = \frac{2 \times PR \times RC}{PR + RC}, \tag{14}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN}, \tag{15}$$

$$E = \frac{(TP + FP) \times (TP + FN) + (FN + TN) \times (FP + TN)}{(TP + TN + FP + FN)^2}, \tag{16}$$

$$KAPPA = \frac{OA - E}{1 - E}, \tag{17}$$

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

In the formulas mentioned above, TP denotes the true positive, which corresponds to the area of change predicted correctly; FP signifies the false positive, representing the unchanged area mistakenly identified as changed; TN is the true negative, referring to the area that is correctly identified as unchanged; FN indicates the false negative, which is the area of change incorrectly labeled as unchanged.

2.3.2. Multi-Scale Deep Supervised Training

To assist the network in better learning useful feature representations and to accelerate the training process, we incorporate additional loss functions at multiple intermediate layers of the network for multi-scale supervision [52,62]. Changes in change detection tasks can occur at various scales, from large-scale geographical or architectural changes to minor object movements or alterations. Multi-scale supervision ensures that the network effectively learns features at all scales, thereby enhancing its ability to detect changes across different scales. The goal of multi-scale supervision training is to minimize the following:

$$\mathcal{L} = \mathcal{L}_1 + \frac{\mathcal{L}_2}{\gamma_2} + \frac{\mathcal{L}_3}{\gamma_3}, \tag{19}$$

$$\gamma_2 = \frac{\mathcal{L}_2}{\mathcal{L}_1}, \tag{20}$$

$$\gamma_3 = \frac{\mathcal{L}_3}{\mathcal{L}_1}. \tag{21}$$

For losses at different levels, we sequentially define them as $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ from shallow to deep layers. γ_2 and γ_3 represent the weights of auxiliary losses at different scales, respectively. We use shallow layer losses as a baseline because shallow layers are usually able to capture more details and texture information, which is crucial for change detection tasks. Moreover, features in shallow layers are closer to the input data, making the loss calculation potentially more stable.

3. Experiment and Results.

In this part, we use the GZ-CD, SYSU-CD, and LEVIR-CD datasets for ablation studies and comparison experiments to thoroughly assess the performance of our proposed network and components.

3.1. Experimental Details

The research was conducted using a GeForce RTX 4090 GPU manufactured by NVIDIA, with the chip produced in Taiwan, China, and employed PyTorch (version 2.1) with CUDA 12.1 as the foundational framework. During training, the batch size was established at six, with an initial learning rate (lr_0) of 0.0001. This learning rate is subject to dynamic modification, adhering to a polynomial adjustment strategy. The computational expression for this adjustment is detailed as follows:

$$lr = lr_0 * \left(1 - \frac{epoch}{num\_epoch}\right)^p. \tag{22}$$

The new learning rate, denoted as lr, is calculated from the initial learning rate lr_0, taking into account the current iteration number ($epoch$), the maximum iteration number (num_epoch), and a constant p that controls the decay rate. The epoch count is fixed at 200. Moreover, the network employs binary cross-entropy loss as its loss function and utilizes the Adam optimization algorithm.

3.2. Ablation Experiments on LEVIR-CD

Selecting the right backbone network is crucial for our experiments. We considered ResNet18, ResNet34, ResNet50, VGG16, and VGG19. Table 1, highlighting the best results in bold, indicates ResNet34's superior performance.

Table 1. Comparative experiments of AMFNet under different backbone networks (best results are highlighted in bold type).

Backbone	PR (%)	RC (%)	IoU (%)	F_1 (%)
VGG19	82.53	82.79	76.16	79.43
VGG16	85.36	84.68	78.54	80.21
ResNet18	93.98	90.38	81.54	89.36
ResNet50	94.22	90.35	82.54	90.05
ResNet34	**95.51**	**92.15**	**83.13**	**90.79**

By adding or removing modules, we examine the function of each component in order to comprehend complicated neural networks. We use the LEVIR-CD dataset for ablation studies to assess these modules' performance on the backbone network. To directly view the effectiveness of the model, we focus on the PR (precision), RC (recall), IoU (ontersection over union), and F_1 metrics. The details are exhibited in Table 2.

Table 2. The assessment of our proposed module through ablative experiments (best results are highlighted in bold type).

Method	PR (%)	RC (%)	IoU (%)	F_1 (%)	Time (ms)
Baseline	94.10	87.72	80.83	88.74	15.12
Baseline + BFAM	94.57	89.31	81.61	89.87	20.47
Baseline + BFAM + BFM	95.16	88.54	82.21	90.24	22.32
Baseline + BFAM + BFM + IAM	95.30	90.99	82.73	90.55	24.23
Baseline + BFAM + BFM + IAM + aux_loss	**95.51**	**92.15**	**83.13**	**90.79**	27.89

1. Ablation experiments of BFAM: We propose the attention-guided BFAM module, which effectively incorporates temporal dynamics into the feature fusion process at

both channel and spatial dimensions. This module enhances IoU and F1 scores by 0.78% and 1.13 %, respectively, validating the effectiveness of the proposed module. Table 3 shows the ablation experiment of the convolution kernel size k used for one-dimensional convolution at the channel latitude. The experiment proves that the model performs best when k is an adaptive channel size. From heatmap2 in Figure 10, it can be observed that, compared to heatmap1, the addition of BFAM significantly reduces the areas of misjudgment. The weights along the edges of the target buildings become more pronounced, leading to a more precise localization of the edges.

2. Ablation experiments of BFM: Our proposed BFM facilitates the network in precisely identifying changing locations during the feature decoding phase by integrating global and differential information, thereby enhancing the representation of texture and edge features. Experimental results shown in Table 2 demonstrate that BFM successfully integrates two semantic pieces of information, improving the MIoU score by 0.60% and the F1 score by 0.37%, enhancing the model's accuracy.
3. Ablation experiments of IAM: The attention module enables the network to adaptively adjust the weights and pixel positions across each channel, emphasizing factors related to changes while suppressing irrelevant ones. This method is a crucial approach to improving feature extraction efficiency in the network. The experimental findings presented in Table 2 show that AMFNet raises the F1 score by 0.31% and the IoU by 0.52%, affirming the accuracy of the proposed module. From Heatmap3 in Figure 10, it is evident that employing interactive features for decoding, a key innovation in this paper, achieves remarkable results. The decoded feature maps undergo further processing through BFM and IAM, leading to a higher emphasis on the target regions. This results in a reduction of misjudgments and omissions along the edges, making the distinction between changed and unchanged regions more apparent.
4. Ablation experiments of multi-scale supervised training: In order to enhance the detection capability of changes at various scales, we add losses from different layers in the decoding stage to the overall training loss in a certain proportion. This results in an improvement of 0.40% and 0.78% in the F1 and IoU scores, respectively, strengthening the model's robustness.

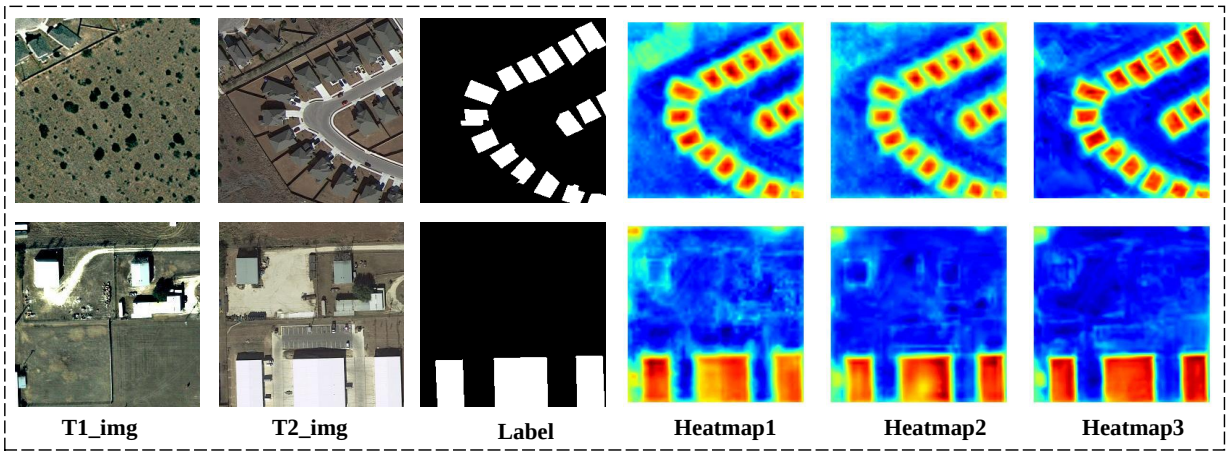

Figure 10. Heatmaps depicting the ablation of different modules. Heatmap1 represents the heatmap of the backbone network, Heatmap2 is the heatmap after adding the BFAM module, and Heatmap3 is the heatmap after adding the BFAM, BFM, and IAM modules.

Table 4 presents the ablation experiment results of BFAM and IAM on the LEVIR-CD dataset. Both modules are attention-guided fusion modules, with BFAM fusing multi-scale bi-temporal feature maps in the encoding stage and IAM highlighting changing regions and suppressing non-changing regions in the decoding stage. For comparison, we selected CBAM [63] as a conventional attention mechanism model. The first and second rows indicate that using BFAM allows concentrating on the more critical aspects between

bi-temporal features in the encoding stage, effectively fusing the bi-temporal features, thereby improving performance. The data in the first and third rows suggest that relative to conventional attention models, employing IAM enhances the merging of multi-semantic features during the decoding phase, thus boosting performance. The fourth row shows that using both BFAM and IAM simultaneously can further enhance performance. According to the data above, change detection capabilities may be further enhanced by integrating BFAM and IAM, which together can boost performance even more.

Table 3. Ablation experiment of one-dimensional convolution kernel of channel latitude in BFAM (best results are highlighted in bold type).

Kernel Size	IoU (%)	F_1 (%)
$k = 3$	82.60	89.46
$k = 5$	82.97	90.30
$k = 7$	81.78	88.35
k is adaptive	**83.13**	**90.79**

Table 4. Attention-guided module ablation experiments (best results are highlighted in bold type).

Encoding Stage Fusion Unit	Decoding Stage Interaction Unit	IoU (%)	F_1 (%)
Add	Regular Attention Module	81.84	89.96
BFAM	Regular Attention Module	82.61	90.30
Add	IAM	82.87	90.12
BFAM	IAM	**83.13**	**90.79**

3.3. Comparative Experiments with Other Classical Networks

3.3.1. Comparative Experiments of Different Algorithms on LEVIR-CD

To fully comprehend and analyze the performance of our model, we contrasted it with many models. Every method applied the same training approach to guarantee comparative fairness, ensuring reliable results. The specific quantified results of various model evaluation metrics are listed in Table 5. Params (M) represents the total number of trainable parameters in the model, expressed in millions, indicating the model's complexity and capacity. FLOPs (G) denotes the number of floating-point operations required to run the model once, measured in billions, reflecting the computational cost. Time (ms) measures the time taken to execute one forward pass of the model in milliseconds, providing a direct metric of the model's inference speed. From the comparative experimental results, it is evident that among the change detection methods using deep learning, FC-EF performs the worst, with F1 and IoU scores of only 83.17% and 71.19%, respectively. In contrast, AMFNet demonstrated increases of 0.69% in F1 score and 1.15% in IoU, alongside notable improvements in precision (+2.98%), recall (+2.39%), Kappa (+0.72%), and overall accuracy (+0.05%). Our approach exceeds other algorithms in terms of overall performance based on many measures, which is a clear indication of its potential superiority in practical applications.

To visualize the prediction results, Figure 11 displays the predicted maps of different algorithms on the LEVIR-CD dataset. We selected three sets of different remote sensing images and generated prediction maps, further illustrating that our algorithm outperforms other state-of-the-art algorithms. Here, white indicates true positives, black indicates true negatives, red indicates false positives, and green indicates false negatives. From the prediction results of the three sets of comparative experiments, it can be observed that other change detection models exhibit more severe omissions and exaggerations, especially in areas with similar features such as color and lighting. In contrast, our proposed algorithm performs better in detecting small-area and high-similarity change regions, providing more accurate predictions of edge details.

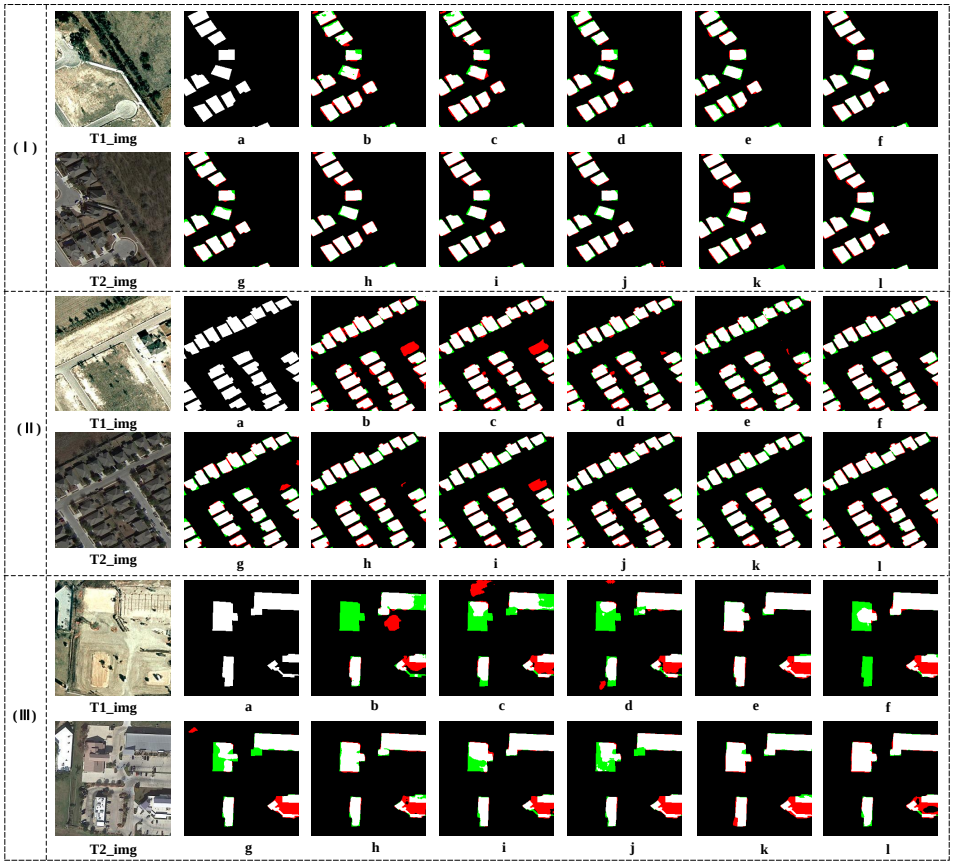

Figure 11. Three groups of comparative diagrams illustrating the performance of different algorithms on the LEVIR-CD dataset. (**a**–**l**) corresponds to the predicted maps of labels, FC-EF, FC-Siam-Conc, FC-Siam-Diff, ChangNet, DSIFN, BIT, ICFNet, SNUNet, DMINet, SAGNet, and our AMFNet network.

Table 5. Comparative experiments on the LEVIR-CD dataset (best results are highlighted in bold type).

Method	PR(%)	RC (%)	OA (%)	Kappa (%)	IoU (%)	F_1 (%)	Params (M)	FLOPs (G)	Time (ms)
FC-EF [64]	85.58	80.89	98.33	82.30	71.19	83.17	1.35	3.57	7.59
FC-Siam-Diff [64]	89.49	80.67	98.53	84.08	73.69	84.85	1.35	4.72	5.13
FC-Siam-Conc [64]	86.76	85.83	98.61	85.56	75.89	86.29	1.55	5.32	5.22
ChangeNet [65]	91.63	86.88	98.93	88.63	80.49	89.19	47.20	10.91	17.01
DSIFN [66]	91.53	85.70	98.87	87.75	79.12	88.34	35.73	82.26	12.13
BIT [41]	91.26	88.51	98.98	89.33	81.59	89.86	3.49	10.63	16.12
ICIFNet [67]	91.31	87.23	98.56	89.16	81.24	89.18	23.84	24.51	49.53
SNUNet [68]	91.51	88.51	99.00	89.46	81.79	89.98	12.03	54.82	9.66
DMINet [69]	92.02	87.77	98.99	89.31	81.56	89.85	6.24	14.55	12.87
SAGNet [55]	91.79	88.76	99.02	89.58	81.98	90.10	32.23	12.25	25.32
Ours	**94.77**	**91.15**	**99.07**	**90.30**	**83.13**	**90.79**	30.27	10.81	27.89

3.3.2. Comparative Experiments of Different Algorithms on GZ-CD

We ran comparison tests on the GZ-CD dataset to confirm the effectiveness of our AMFNet method. The outcomes are shown in Table 6. It is evident from the F1 and IoU scores that FC-Siam-Diff performs the worst, with F1 and IoU scores of only 71.06% and 55.11%, respectively. The remaining change detection algorithms show improvements in all four metrics, with our AMFNet exhibiting the best performance. F1 and IoU scores reached 88.09% and 69.85%, respectively, which are 1.79% and 2.80% higher than the second-ranked SAGNet, indicating excellent generalization ability and robustness of our algorithm.

Table 6. Comparative experiments on the GZ-CD dataset (best results are highlighted in bold type).

Method	PR (%)	RC (%)	OA (%)	Kappa (%)	IoU (%)	F_1 (%)
FC-EF [64]	79.86	65.53	95.28	69.44	56.24	71.99
FC-Siam-Diff [64]	82.70	57.99	94.99	65.55	51.72	68.18
FC-Siam-Conc [64]	82.16	62.80	95.29	68.67	55.26	71.19
ChangeNet [65]	88.63	82.99	97.44	84.32	75.01	85.72
DSIFN [66]	89.35	75.46	96.91	79.83	68.76	81.49
BIT [41]	86.80	82.04	97.18	82.80	72.94	84.35
ICIFNet [67]	88.09	81.31	97.25	83.05	73.25	84.56
SNUNet [68]	89.00	84.80	97.62	85.54	76.75	86.85
DMINet [69]	86.62	82.85	97.23	83.17	73.45	84.70
SAGNet [55]	89.56	84.05	97.58	84.98	75.91	86.30
Ours	**90.40**	**89.74**	**97.85**	**86.91**	**78.71**	**88.09**

On the GZ-CD dataset, the predicted maps for each comparative experiment are illustrated in Figure 12. Other change detection algorithms often exhibit extensive misjudgments and omissions at the edges of change regions. In contrast, our proposed algorithm achieves more accurate edge detection and demonstrates stronger detection capabilities for small targets similar to the background. In Figure 12(II), the portion enclosed by the three buildings on the right side of the image remains unchanged and is still land. However, due to external environmental changes such as lighting and seasonality, many algorithms misclassify it as a changed area. Our algorithm accurately identifies this as an unchanged area, proving the effectiveness of our AMFNet algorithm in change detection tasks.

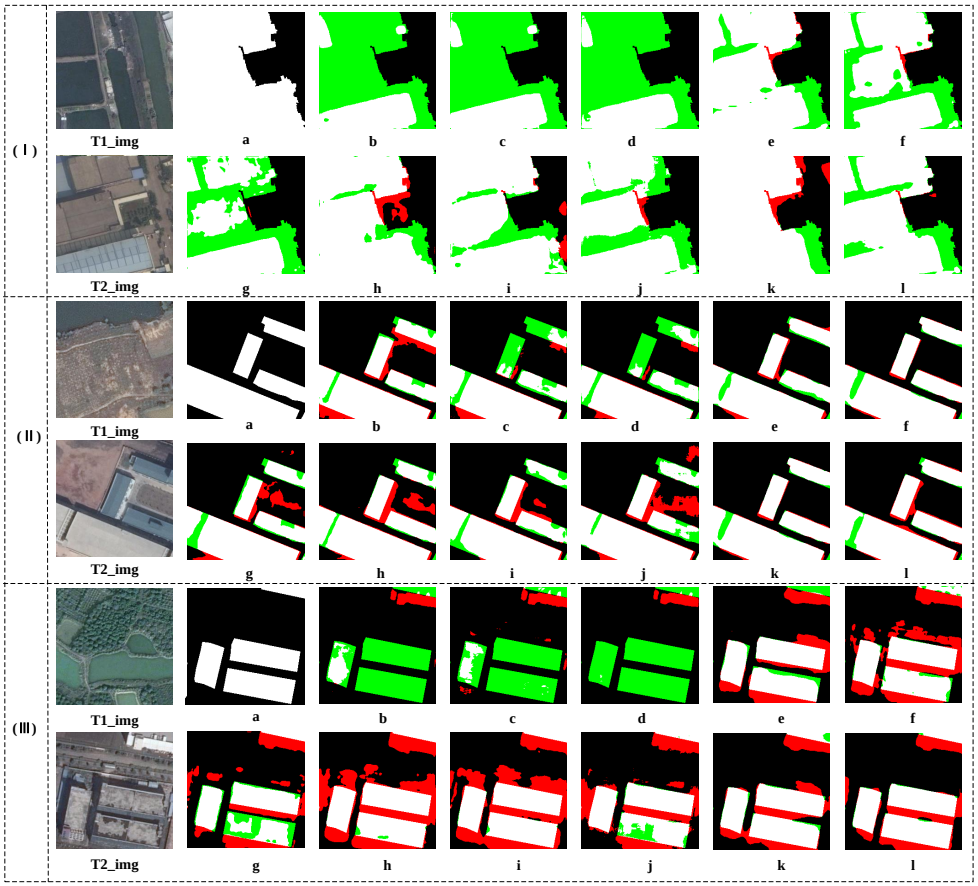

Figure 12. Three groups of comparison diagrams of different algorithms on GZ-CD dataset. (**a**–**l**) corresponds to the predicted maps of labels, FC-EF, FC-Siam-Conc, FC-Siam-Diff, ChangNet, DSIFN, BIT, ICFNet, SNUNet, DMINet, SAGNet, and our AMFNet network.

3.3.3. Comparative Experiments of Different Algorithms on SYSU-CD

Concerning the SYSU-CD dataset, our model AMFNet achieved the highest IoU and F1 indicators, reaching 69.85% and 82.25%, respectively. The outcomes are shown in Table 7.

Table 7. Comparative experiments on the SYSU-CD dataset (best results are highlighted in bold type).

Method	PR (%)	RC(%)	OA (%)	Kappa (%)	IoU (%)	F_1 (%)
FC-EF [64]	78.78	76.69	89.63	70.97	63.56	77.72
FC-Siam-Diff [64]	80.35	74.26	88.71	64.42	55.11	71.06
FC-Siam-Conc[64]	81.51	75.11	90.11	71.80	64.17	78.18
ChangeNet [65]	79.91	71.11	88.97	68.19	60.33	75.25
DSIFN [66]	78.82	81.30	90.44	73.76	66.72	80.04
BIT [41]	81.22	73.87	89.81	70.81	63.09	77.37
ICIFNet [67]	78.23	74.38	89.08	69.17	61.62	76.25
SNUNet [68]	79.37	78.39	90.10	72.42	65.13	78.88
DMINet [69]	81.54	79.44	91.15	74.59	67.06	80.28
SAGNet [55]	81.25	81.76	91.72	76.57	69.31	81.87
Ours	**88.23**	**82.51**	**92.30**	**77.38**	**69.85**	**82.25**

Figure 13 illustrates the prediction results of different algorithms on the SYSU-CD dataset. Notably, in Figure 13(II), the bottom-right corner contains a shadow, and the outlines of the buildings are somewhat blurred, posing a significant challenge to the change detection task. Other change detection algorithms perform poorly in this region, resulting in numerous misjudgments. In contrast, our proposed algorithm accurately identifies changes in this part. From Figure 13, it is evident that the AMFNet algorithm achieves more precise edge detection and can effectively extract image features for accurate judgment even in challenging environments. It demonstrates robustness in situations where the target area is highly similar to the background.

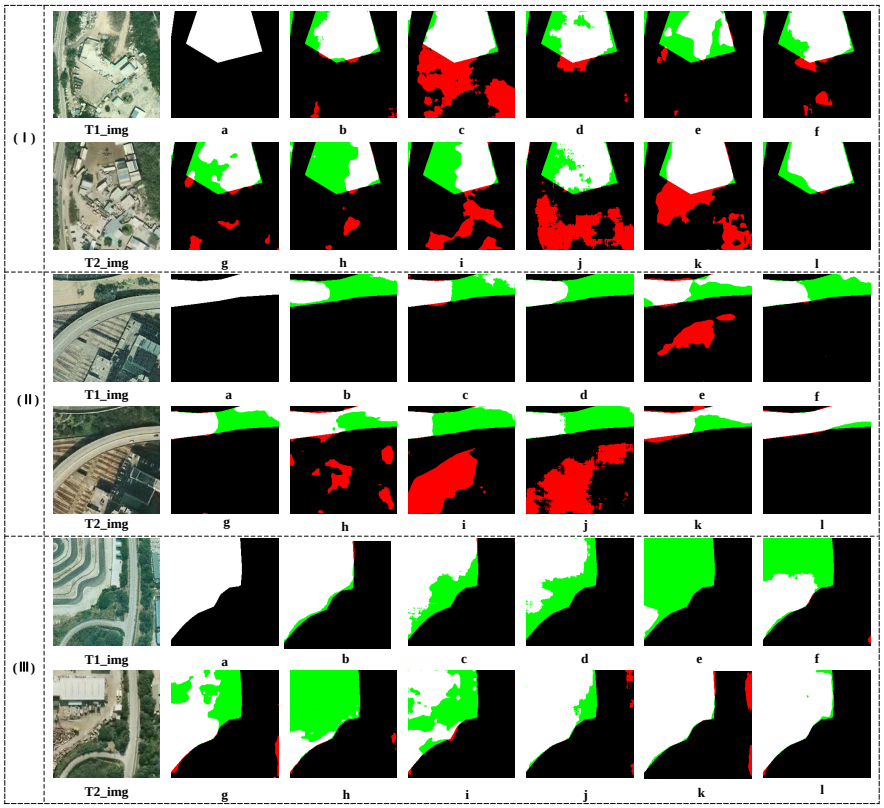

Figure 13. Three groups of comparative diagrams illustrating the performance of different algorithms on the SYSU-CD dataset. (**a**–**l**) corresponds to the predicted maps of labels, FC-EF, FC-Siam-Conc, FC-Siam-Diff, ChangNet, DSIFN, BIT, ICFNet, SNUNet, DMINet, SAGNet, and our AMFNet network.

4. Discussion

4.1. Advantages of the Proposed Method

The proposed attention-guided multi-scale fusion network (AMFNet) introduces significant advancements in the field of change detection. By integrating the bi-temporal feature attention module (BFAM), AMFNet effectively manages the interaction of bi-temporal features, reducing redundancy and enhancing the perception of changes over time. This is a crucial improvement over traditional methods that rely merely on simple arithmetic operations. The incorporation of the bi-temporal fusion module (BFM) in the decoder further refines the process by selectively fusing semantic information, which significantly reduces the intrusion of irrelevant data. Furthermore, the innovative use of the interactive attention module (IAM) in the decoding process not only preserves the temporal sequence but also emphasizes the importance of non-local relationships, allowing for a more nuanced understanding of the scene dynamics. These methodological enhancements enable AMFNet to achieve superior performance metrics, as evidenced by its high F1 scores and intersection over union (IoU) metrics across multiple datasets.

4.2. Limitations and Expectations

While the proposed AMFNet demonstrates impressive performance, it is not without its limitations. The complexity introduced by attention mechanisms and multiscale fusion could lead to increased computational demands [70], potentially hindering real-time processing capabilities or deployment in computationally constrained environments [71]. The model's sensitivity to hyperparameters and network architecture choices also poses a challenge, necessitating careful tuning to achieve the best possible performance. Furthermore, the current evaluation of the model's effectiveness has been limited to specific datasets, and its adaptability to larger-scale remote sensing images with more diverse regions [69] remains uncertain. Looking ahead, it is essential to concentrate research efforts on enhancing computational efficiency, reducing the model's sensitivity to parameter settings, and improving scalability. Such advancements will be crucial in extending the applicability of AMFNet to a wider array of real-world scenarios, thereby solidifying its position as a robust tool in the realm of remote sensing and change detection.

5. Conclusions

In this paper, we proposed an attention-guided multiscale fusion network. The traditional change detection algorithm simply adds and subtracts bi-temporal features, which have a large amount of redundant information and insufficient perception of the tense. So, in the encoding part, we utilized BFAM to facilitate the interaction of bi-temporal features at the same level, eliminating redundant information. Our decoding process was based on interactive feature maps, preserving both the temporal sequence and rich non-local relationships of bi-temporal features during further fusion. By incorporating BFM into the decoder network to fuse global semantic and differential semantic information, we avoided introducing a significant amount of irrelevant semantic information through simple fusion methods. By assigning weights adaptively to different regions, the IAM sharpened the network's emphasis on differentiating targets from backgrounds and enhances its capacity to identify tiny objects and edge features. Our proposed AMFNet outperforms existing algorithms on the LEVIR-CD, GZ-CD, and SYSU-CD datasets, according to experimental findings. Specifically, it achieved F1 scores of 90.79%, 88.09%, and 82.25%, and IoU metrics of 83.13%, 78.71%, and 69.85% for each dataset, respectively. The model exhibited excellent generalization and robustness.

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

Funding: This work was supported, in part, by the National Natural Science Foundation of PR China (42075130).

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

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

References

1. Amare, M.T.; Demissie, S.T.; Beza, S.A.; Erena, S.H. Land cover change detection and prediction in the Fafan catchment of Ethiopia. *J. Geovis. Spat. Anal.* **2023**, *7*, 19. [CrossRef]
2. Eisavi, V.; Homayouni, S.; Karami, J. Integration of remotely sensed spatial and spectral information for change detection using FAHP. *J. Fac. For. Istanb. Univ.* **2016**, *66*, 524–538. [CrossRef]
3. de Alwis Pitts, D.A.; So, E. Enhanced change detection index for disaster response, recovery assessment and monitoring of accessibility and open spaces (camp sites). *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *57*, 49–60. [CrossRef]
4. Tan, K.; Ma, W.; Chen, L.; Wang, H.; Du, Q.; Du, P.; Yan, B.; Liu, R.; Li, H. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. *J. Hazard. Mater.* **2021**, *401*, 123288. [CrossRef]
5. Qin, H.; Wang, J.; Mao, X.; Zhao, Z.; Gao, X.; Lu, W. An Improved Faster R-CNN Method for Landslide Detection in Remote Sensing Images. *J. Geovis. Spat. Anal.* **2024**, *8*, 2. [CrossRef]
6. Ji, R.; Tan, K.; Wang, X.; Pan, C.; Xin, L. Spatiotemporal monitoring of a grassland ecosystem and its net primary production using Google Earth Engine: A case study of inner mongolia from 2000 to 2020. *Remote Sens.* **2021**, *13*, 4480. [CrossRef]
7. Kokila, S.; Jayachandran, A. Hybrid Behrens-Fisher-and gray contrast–based feature point selection for building detection from satellite images. *J. Geovis. Spat. Anal.* **2023**, *7*, 8. [CrossRef]
8. 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]
9. Weismiller, R.; Kristof, S.; Scholz, D.; Anuta, P.; Momin, S. Change detection in coastal zone environments. *Photogramm. Eng. Remote Sens.* **1977**, *43*, 1533–1539.
10. Ke, L.; Lin, Y.; Zeng, Z.; Zhang, L.; Meng, L. Adaptive Change Detection With Significance Test. *IEEE Access* **2018**, *6*, 27442–27450. [CrossRef]
11. Rignot, E.J.; Van Zyl, J.J. Change detection techniques for ERS-1 SAR data. *IEEE Trans. Geosci. Remote Sens.* **1993**, *31*, 896–906. [CrossRef]
12. Ridd, M.K.; Liu, J. A comparison of four algorithms for change detection in an urban environment. *Remote Sens. Environ.* **1998**, *63*, 95–100. [CrossRef]
13. Ferraris, V.; Dobigeon, N.; Wei, Q.; Chabert, M. Detecting changes between optical images of different spatial and spectral resolutions: A fusion-based approach. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 1566–1578. [CrossRef]
14. Du, P.; Wang, X.; Chen, D.; Liu, S.; Lin, C.; Meng, Y. An improved change detection approach using tri-temporal logic-verified change vector analysis. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 278–293. [CrossRef]
15. Deng, J.; Wang, K.; Deng, Y.; Qi, G. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *Int. J. Remote Sens.* **2008**, *29*, 4823–4838. [CrossRef]
16. Zhang, X.; Yang, P.; Zhou, M. Multireceiver SAS imagery with generalized PCA. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1502205. [CrossRef]
17. Raj, J.R.; Srinivasulu, S. Change detection of images based on multivariate alteration detection method. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 847–850.
18. Chen, H.; Yokoya, N.; Chini, M. Fourier domain structural relationship analysis for unsupervised multimodal change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 99–114. [CrossRef]
19. Ji, L.; Zhao, J.; Zhao, Z. A Novel End-to-End Unsupervised Change Detection Method with Self-Adaptive Superpixel Segmentation for SAR Images. *Remote Sens.* **2023**, *15*, 1724. [CrossRef]
20. Dou, P.; Han, Z. Quantifying Land Use/Land Cover Change and Urban Expansion in Dongguan, China, From 1987 to 2020. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 201–209. [CrossRef]
21. Dou, P.; Chen, Y. Dynamic monitoring of land-use/land-cover change and urban expansion in Shenzhen using Landsat imagery from 1988 to 2015. *Int. J. Remote Sens.* **2017**, *38*, 5388–5407. [CrossRef]
22. Juan, S.; Gui-Jin, W.; Xing-Gang, L.; Dai-Zhi, L. A change detection algorithm for man-made objects based on multi-temporal remote sensing images. *Acta Autom. Sin.* **2008**, *34*, 1040–1046.
23. Wan, L.; Xiang, Y.; You, H. A post-classification comparison method for SAR and optical images change detection. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1026–1030. [CrossRef]
24. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.

25. Wu, C.; Du, B.; Cui, X.; Zhang, L. A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion. *Remote Sens. Environ.* **2017**, *199*, 241–255. [CrossRef]
26. Tan, K.; Zhang, Y.; Wang, X.; Chen, Y. Object-based change detection using multiple classifiers and multi-scale uncertainty analysis. *Remote Sens.* **2019**, *11*, 359. [CrossRef]
27. Han, Y.; Javed, A.; Jung, S.; Liu, S. Object-based change detection of very high resolution images by fusing pixel-based change detection results using weighted Dempster–Shafer theory. *Remote Sens.* **2020**, *12*, 983. [CrossRef]
28. Dou, P.; Huang, C.; Han, W.; Hou, J.; Zhang, Y.; Gu, J. Remote sensing image classification using an ensemble framework without multiple classifiers. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 190–209. [CrossRef]
29. Jiang, S.; Dong, R.; Wang, J.; Xia, M. Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network. *Systems* **2023**, *11*, 305. [CrossRef]
30. Dou, P.; Shen, H.; Li, Z.; Guan, X. Time series remote sensing image classification framework using combination of deep learning and multiple classifiers system. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102477. [CrossRef]
31. 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]
32. Wang, X.; Yan, X.; Tan, K.; Pan, C.; Ding, J.; Liu, Z.; Dong, X. Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *122*, 103456. [CrossRef]
33. 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]
34. 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]
35. Weng, L.; Pang, K.; Xia, M.; Lin, H.; Qian, M.; Zhu, C. Sgformer: A Local and Global Features Coupling Network for Semantic Segmentation of Land Cover. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6812–6824. [CrossRef]
36. Niu, C.; Tan, K.; Jia, X.; Wang, X. Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery. *Environ. Pollut.* **2021**, *286*, 117534. [CrossRef] [PubMed]
37. 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]
38. 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.
39. Li, X.; He, M.; Li, H.; Shen, H. A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8017505. [CrossRef]
40. 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.
41. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5900318. [CrossRef]
42. Xing, Y.; Jiang, J.; Xiang, J.; Yan, E.; Song, Y.; Mo, D. LightCDNet: Lightweight Change Detection Network Based on VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2504105. [CrossRef]
43. 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]
44. Lv, Z.; Huang, H.; Sun, W.; Lei, T.; Benediktsson, J.A.; Li, J. Novel enhanced UNet for change detection using multimodal remote sensing image. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2505405. [CrossRef]
45. Lv, Z.; Liu, J.; Sun, W.; Lei, T.; Benediktsson, J.A.; Jia, X. Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4411115. [CrossRef]
46. Lv, Z.; Zhang, P.; Sun, W.; Lei, T.; Benediktsson, J.A.; Li, P. Sample Iterative Enhancement Approach for Improving Classification Performance of Hyperspectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2023**, *21*, 2500605. [CrossRef]
47. 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]
48. Song, L.; Xia, M.; Jin, J.; Qian, M.; Zhang, Y. SUACDNet: Attentional change detection network based on siamese U-shaped structure. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102597. [CrossRef]
49. Luo, F.; Zhou, T.; Liu, J.; Guo, T.; Gong, X.; Ren, J. Multiscale diff-changed feature fusion network for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5502713. [CrossRef]
50. 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. [CrossRef]
51. 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.* **2022**, *16*, 32–43. [CrossRef]
52. 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 June 2017; pp. 2881–2890.

53. 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.
54. Deepanshi, R.; Barkur, D.; Suresh, S.; Lal, C. S.; Reddy, P. G.; 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.* **2022**, *7*, 537–551. [CrossRef]
55. 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]
56. 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]
57. Zhao, S.; Zhang, X.; Xiao, P.; He, G. Exchanging Dual-Encoder–Decoder: A New Strategy for Change Detection With Semantic Guidance and Spatial Localization. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4508016. [CrossRef]
58. 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.
59. Zhang, H.; Chen, H.; Zhou, C.; Chen, K.; Liu, C.; Zou, Z.; Shi, Z. BiFA: Remote Sensing Image Change Detection with Bitemporal Feature Alignment. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5614317. [CrossRef]
60. 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.* **2021**, *60*, 5604816. [CrossRef]
61. Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; Ding, H.; Huang, X. SemiCDNet: A Semisupervised Convolutional Neural Network for Change Detection in High Resolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5891–5906. [CrossRef]
62. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In Proceedings of the Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
63. Woo, S.; Park, J.; Lee, J.Y.; 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.
64. 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.
65. 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.
66. 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]
67. 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]
68. 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*, 8007805. [CrossRef]
69. Feng, Y.; Jiang, J.; Xu, H.; Zheng, J. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4401015. [CrossRef]
70. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]
71. Chen, K.; Liu, C.; Li, W.; Liu, Z.; Chen, H.; Zhang, H.; Zou, Z.; Shi, Z. Time Travelling Pixels: Bitemporal Features Integration with Foundation Model for Remote Sensing Image Change Detection. *arXiv* **2023**, arXiv:2312.16202.