remote sensing

MDPI

Article

MDANet: A High-Resolution City Change Detection Network Based on Difference and Attention Mechanisms under Multi-Scale Feature Fusion

Shanshan Jiang [1], Haifeng Lin [2], Hongjin Ren [3], Ziwei Hu [3], Liguo Weng [3] and Min Xia [3,4,*]

[1] School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; jss@nuist.edu.cn
[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; haifeng.lin@njfu.edu.cn
[3] Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212220006@nuist.edu.cn (H.R.); 202212490021@nuist.edu.cn (Z.H.); 002311@nuist.edu.cn (L.W.)
[4] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: xiamin@nuist.edu.cn

Abstract: In the domains of geographic information systems and remote sensing image analysis, change detection is vital for examining surface variations in high-resolution remote sensing pictures. However, the intricate texture characteristics and rich details found in high-resolution remote sensing photos are difficult for conventional change detection systems to deal with. Target misdetection, missed detections, and edge blurring are further problems with current deep learning-based methods. This research proposes a high-resolution city change detection network based on difference and attention mechanisms under multi-scale feature fusion (MDANet) to address these issues and improve the accuracy of change detection. First, to extract features from dual-temporal remote sensing pictures, we use the Siamese architecture as the encoder network. The Difference Feature Module (DFM) is employed to learn the difference information between the dual-temporal remote sensing images. Second, the extracted difference features are optimized with the Attention Refinement Module (ARM). The Cross-Scale Fusion Module (CSFM) combines and enhances the optimized attention features, effectively capturing subtle differences in remote sensing images and learning the finer details of change targets. Finally, thorough tests on the BTCDD dataset, LEVIR-CD dataset, and CDD dataset show that the MDANet algorithm performs at a cutting-edge level.

Keywords: change detection; attention mechanism; remote sensing images; Siamese structure; MDANet

check for updates

Citation: Jiang, S.; Lin, H.; Ren, H.; Hu, Z.; Weng, L.; Xia, M. MDANet: A High-Resolution City Change Detection Network Based on Difference and Attention Mechanisms under Multi-Scale Feature Fusion. *Remote Sens.* **2022**, *16*, 1387. https://doi.org/10.3390/rs16081387

Academic Editor: Farid Melgani

Received: 19 February 2024
Revised: 28 March 2024
Accepted: 3 April 2024
Published: 14 April 2024

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the continuous development of society and the advancement of science and technology, human activities have had an increasingly significant impact on the natural environment. Activities such as deforestation, land reclamation, and the excessive exploitation and irrational utilization of resources have brought a series of potential hazards to the natural environment [1–3]. At the same time, the natural environment itself undergoes constant changes, with natural disasters like earthquakes, floods, tsunamis, and landslides posing threats to human life and well-being. Therefore, different Earth-observing satellites are used to provide in-the-moment observations and detailed analyses of surface conditions through remote sensing imaging technology in order to better monitor and analyze changes on the Earth's surface and achieve a harmonious coexistence between humanity and nature. With its intuitive representation of the distribution of characteristics and information, remote sensing pictures have emerged as an essential tool for examining Earth's resources and environmental changes. Change detection in remote sensing pictures

is a crucial application are; it evaluates dual-temporal or multi-temporal photos of the same geographic area to spot changes on the Earth's surface [4,5]. Its primary objective is to detect areas of interest that have undergone changes while excluding unrelated changes caused by factors such as lighting and shadows [6]. With the rapid advancement of remote sensing technology, change detection plays a significant role in various fields, including economic development [7,8], environmental monitoring [9,10], and national defense [11]. The research in change detection aims to effectively capture the evolution of the Earth's surface, helping us better understand and respond to the ever-evolving environmental and societal needs.

Scholars from both home and abroad have undertaken in-depth study in the area of change detection and proposed a range of change detection techniques over the years. These methods can be summarized and categorized from different dimensions. First, they may be separated into supervised change detection and unsupervised change detection depending on whether they require human intervention [12]. Traditional unsupervised change detection methods include multivariate alteration detection (MAD) [13] and iteratively reweighted multivariate alteration detection (IR-MAD) [14]. The essence of MAD is a typical correlation analysis in multivariate statistical analysis, but this algorithm cannot handle multi-element remote sensing images better. Therefore, the IR-MAD algorithm is proposed based on MAD. The core idea of the IR-MAD algorithm is to assign new weights to the pixels in the two images in an iterative manner. Unchanged pixels have larger weights, and the weight value of a pixel is the main basis for determining whether the pixel has changed. The iteration stops when the weight of each pixel gradually stabilizes until it becomes unchanged. The final weight of each pixel is compared to a threshold to determine whether it belongs to a changed or unchanged pixel. Finally, the unchanged pixels are extracted as feature pixels. Different from unsupervised learning, supervised learning learns which areas have changed through labeled training sets. Commonly used methods include random forest [15], support vector machine [16], etc. Volpi et al. [17] extracted texture features through a gray-level co-occurrence matrix, and used support vector machine as a classification algorithm to obtain the change area and change type. Wang et al. [18] used K nearest neighbor, random forest and support vector machine for model integration, and achieved better detection results than the change detection method that only uses a single machine learning model. Secondly, they can be categorized based on the fundamental processing unit into pixel-level, object-level, and scene-level methods. The pixel-based change detection method uses pixels as the basic research unit and obtains change maps by comparing pixels in multi-temporal remote sensing images one by one. Gong et al. [19] proposed a neighborhood-based ratio method to generate difference maps. Although this method is superior to traditional difference map generation methods, the change detection results are seriously affected by noise. Object-based change detection methods first segment images into small homogeneous objects, which then serve as subsequent classification and building blocks of larger entities for the study of spatial, textural, contextual, geometric, and spectral features. Liu et al. [20] fused spectral features and shape features to improve the discriminability between different ground objects in ultra-high-resolution images to achieve the change detection of buildings. Desclée et al. [21] adopted an object-based and statistical approach to depict multi-date objects through regional merging technology, and successfully identified forest cover changes. Based on scene change detection methods, the image is processed as a whole, rather than at the pixel or object level. It usually utilizes the global features or contextual information of the image to detect changes, such as image histogram, global statistical features, spatial structure, etc. Liu et al. [22] proposed a symmetric convolutional coupling network to convert the local features of heterogeneous images into the same feature space for comparison, and used the probability map matrix to optimize and train the neural network. However, the network has fewer convolutional layers and has poor ability to learn the global features of the image. Finally, depending on whether they make use of neural network models, change detection techniques may be divided into classical techniques and deep learning techniques. The direct

comparison method, which uses methods like differencing [23], ratioing [24], and change vector analysis (CVA) [25], is one of the most used methodologies in conventional remote sensing image change detection methods. Change vector analysis combines image pixel values at two points in time into vector form and detects changes by calculating the difference between these vectors. Chen and Chen [26] calculated the change intensities of texture features and spectral features through CVA, and then used adaptive weighting to fuse the change intensities; Luppino et al. [27] proposed using a cycle-consistent generative adversarial network to transcode images from different sensors into the same domain in an unsupervised manner, and further implemented change detection through CVA. These methods are characterized by their simplicity, typically involving pixel-wise algebraic computations performed after the image preprocessing steps, followed by the use of appropriate thresholds to distinguish between changed and unchanged regions. This may result in insensitivity to changes in different scenarios (such as natural environments, urban environments, and industrial environments), making it difficult to adapt to the change detection requirements of various environments.

With the swift advancement of satellite remote sensing technology, remote sensing images' quality and scope continue to advance, encompassing more extensive data and potential applications [28]. Modern remote sensing applications place a high demand on change detection techniques, making it difficult for existing approaches to keep up, creating new problems for change detection researchers. In order to meet the changing demands in the field of remote sensing, new techniques and technologies must not only increase detection accuracy but also eliminate interference from outside sources. Change detection in remote sensing photos has benefited from deep learning's substantial technological assistance [29]. Automatic image segmentation is made possible by deep learning-based change detection approaches that automatically extract deep features from dual-temporal and multi-temporal remote sensing pictures. This greatly reduces the requirement for human feature engineering. These methods also exhibit greater robustness in handling large-scale change detection tasks [30]. Ding et al. [31] proposed a landslide detection method that combines convolutional neural networks (CNNs) with texture change detection. The possible landslide locations are identified using CNN, and the features are then subjected to texture analysis to identify changes related to landslides. However, the adaptability of this method to different types of terrain and texture changes needs further study. Hou et al. [32] proposed a method that integrates low-rank changes and deep feature changes detection, using fine-tuned VGGNet [33] to extract object-level features. This approach enhances change detection performance by fusing multi-level features. However, low-rank decomposition may not effectively capture complex changes in images, particularly leading to missed detections in scenarios with numerous subtle changes. Wang et al. [34] proposed an end-to-end dual-channel CNN framework suitable for change detection in hyperspectral images. They introduced a hybrid affinity matrix to exploit multi-channel gradient features and aggregated multisource information to enhance detection accuracy. However, there may be a risk of information loss during the extraction of the hybrid affinity matrix and fusion of multisource information. Zhang and Lu [35] proposed a spectral–spatial joint learning network, initially using a Siamese CNN to extract joint representations of spectral and spatial information, followed by feature fusion to enhance the discriminative ability of change detection. Chen and Shi [36] proposed a spatiotemporal attention neural network based on Siamese architecture, which utilizes spatiotemporal attention modules to compute pixel attention weights for different times and locations, generating more discriminative features. The model emphasizes spatial feature extraction, showing significant effectiveness in scenes sensitive to surface changes, particularly performing well in datasets with relatively singular change targets, such as the LEVIR-CD dataset. However, it exhibits slight limitations when handling datasets containing various land covers and seasonal vegetation changes. Chen et al. [37] proposed a dual-attention fully convolutional Siamese network to address the issue of insufficient resistance to pseudo-changes in existing methods. Through a dual-attention mechanism, this approach can capture long-range dependencies, obtain-

ing more discriminative feature representations and thus improving model recognition performance. The method excels in handling pseudo-changes, such as in the CDD dataset, but exhibits slight limitations in handling small targets and edge details. Chen et al. [38] proposed a novel change detection model called BIT based on Transformer. This innovative approach utilizes Transformer modules to enhance the model's capability in extracting spatiotemporal contextual information, thus aiding the model in more effectively identifying regions of interest for change. Chen et al. [39] proposed a feature-constrained change detection network that achieves feature extraction and fusion through dual-temporal features. Additionally, a non-local feature pyramid module is designed at the core of the backbone network, along with a feature fusion module based on dense connections, enabling the extraction and fusion of multi-scale features. Shen et al. [40] proposed a semantic feature-constrained change detection network, which achieves simultaneous operations of feature extraction, semantic segmentation, and change detection. Liao et al. [41] proposed a contrastive learning approach, which enhances the discriminative features of buildings and non-buildings and injects building semantics into the change detection channel. Moreover, to address the issue of inconsistency between historical building polygons and the latest images, deformable convolutional neural networks are employed to learn offsets. The detection of changes in high-resolution remote sensing images is crucial for exploring surface changes and holds significant importance in the fields of geographic information systems and remote sensing image analysis. However, traditional change detection methods often struggle to handle the rich details and complex texture features present in high-resolution remote sensing images. Additionally, deep learning-based algorithms also encounter challenges such as false detections, missed detections, and blurred edges when faced with these complexities. Therefore, there is a need to propose a new change detection method that can effectively address the complexity of high-resolution remote sensing images and improve the accuracy of change detection.

2. Related Work

The fully convolutional neural network (FCN) was initially introduced by Long et al. [42] for image segmentation tasks [43]. By replacing fully connected layers with upsampling layers, it can restore downsampled features to the original size of the input image, achieving end-to-end classification. Based on the manner of processing dual-temporal images, it is divided into single-input networks and Siamese networks. The semantic segmentation networks used for tasks like target extraction and picture classification are analogous to the single-input network structure. It only accepts input from multi-channel photos of a single scenario. In order to satisfy the network's input criteria, the dual-temporal pictures must be combined into a multi-channel image before being fed into the network. Alcantarilla et al. [44] concatenated two dual-temporal images with three channels each into a single image with six channels. Subsequently, this six-channel image is fed into the FCN to perform change detection tasks. Peng et al. [45] combined dual-temporal remote sensing images into one input and then fed them into the UNet++ network for change detection. The dual-temporal picture channels are immediately combined into the network using these models. Although they function with the majority of semantic segmentation networks, they are unable to give deep features for reconstructing specific pictures, which results in channel clutter and poor compactness in the images during the feature extraction step. Siamese networks are now widely used for change detection jobs as a result. The properties of change detection in dual-temporal remote sensing pictures are taken into consideration by Siamese networks. They use an architecture in which a connected structure is formed by two identical neural subnetworks. Siamese networks send the dual-temporal pictures individually into two identical and weight-sharing independent neural networks rather than channel merging them before feeding them into the network. These subnetworks extract the same characteristics via weight sharing [46]. Daudt et al. [46] proposed the use of Siamese networks for change detection and compared it with the single-input network method, demonstrating the effectiveness of Siamese networks. Building

upon this, Guo et al. [47] utilized threshold contrastive loss to modify the fully convolutional Siamese network, reducing the distance between unchanged feature pairs and expanding the distance between changed feature pairs, thereby enhancing performance.

Thanks to developments in Earth observation technology, we can now easily gather a sizable number of high-resolution remote sensing photos, which sets new demands on change detection in remote sensing images. Despite the emergence of numerous change detection algorithms over the past few decades, and the surge in big data and artificial intelligence after 2010 which propelled a new wave of interest in change detection [48,49], driving rapid advancements in the field, existing methods still face challenges given the complexity of change detection tasks in remote sensing imagery. Firstly, in the imaging methods of multi-temporal remote sensing images, factors such as sensor resolution, optical system performance, and image acquisition parameters may have an impact, while imaging limitations may involve environmental conditions, shooting angles, and characteristics of the target surface. These factors can affect the clarity of the final predicted image and the accuracy of its edges. With the continuous improvement of remote sensing image resolution, these issues become more severe, posing challenges to previous change detection methods. Second, deep neural networks are frequently used to implicitly extract picture difference maps in change detection approaches based on semantic segmentation that generally overlay two temporal remote sensing images in the channel dimension. The distinguishing traits might not be adequately captured by this method. Finally, existing methods often focus on integrating multi-scale information or semantic features at the end of the network, while the utilization of rich feature information in the intermediate layers is less emphasized. However, the intermediate layer features may play an important role in supervising model training. Fully exploiting the intermediate layer features may help reduce issues such as false detections, missed detections, and blurred edges during the prediction process. In response to these challenges, we propose a high-resolution remote sensing image change detection network based on multi-scale feature fusion based on difference and attention mechanisms (MDANet). Our approach consists of the following key components: In order to extract distinct land characteristics from two temporal photos, we first employ the Siamese network as the backbone network. Unlike the brute-force fusion of multi-temporal images using semantic segmentation-based methods, this approach selectively extracts remote sensing features from different time phases. Secondly, we introduce a Difference Feature Module (DFM) between the Siamese networks to extract the difference information between dual-temporal remote sensing features. We propose a novel Attention Refinement Module (ARM) between the Siamese networks. The module possesses the capability to adaptively focus on changing regions, thereby mitigating the influence of factors such as lighting variations, seasonal changes, and misregistration errors. Additionally, a Cross-Scale Fusion Module (CSFM) is introduced to maximize the use of differences refined through attention at various hierarchical levels. In order to recover as much of the edge information of change areas as possible, it also takes into account the category of the pixel and concentrates on the semantic integrity of the picture. To sum up, the main contributions we made are as follows:

1. We suggest a brand-new change detection network, named the multi-scale feature fusion-based high-resolution remote sensing image change detection network (MDANet) based on difference and attention mechanisms. The Siamese network is used to extract features from dual-temporal remote sensing pictures individually during the feature decoding phase. Our suggested auxiliary modules are then used to extract change information from these images. In the feature encoding phase, we utilize a residual structure augmented with stripe convolution to restore the change regions of the original image. Stripe convolution emphasizes edge detail features during the restoration process, significantly enhancing the detection performance.
2. Additionally, we have innovatively designed three auxiliary modules, namely the Difference Feature Module (DFM), Attention Refinement Module (ARM), and Cross-Scale Fusion Module (CSFM). DFM conducts difference operations on features extracted

by the Siamese network to highlight change characteristics. Along with eliminating non-change features and adaptively collecting change information, ARM further refines the extracted difference features in both the spatial and the channel dimensions. CSFM effectively integrates change features from different scales, enhancing the model's perception and utilization of features from various scales, reducing the model's dependency on specific information, and improving its generalization ability.

The rest of the paper is as follows: Firstly, a comprehensive exposition of the methodology is provided, detailing the approach we proposed. Subsequently, the datasets used are introduced, offering readers information on their sources and backgrounds. Then, specific details of the experiments are presented, including experimental setups, parameter selections, and methods of performance evaluation. Finally, the work presented in this paper is discussed and summarized.

3. Materials and Methods

3.1. Proposed Approach

This paper introduces a high-resolution city change detection network based on difference and attention mechanisms under multi-scale feature fusion (MDANet). Its aim is to achieve the prediction of change maps with structural integrity and fine boundaries. We will begin by outlining the main network architecture and then provide a detailed explanation of the specific implementation of each module.

3.1.1. Network Architecture

The MDANet encoder–decoder structure is built in this article. The backbone network and three auxiliary modules make up the decoder. Similar to other remote sensing image change detection networks, we have chosen ResNet34 [50] as our encoding network. To extract spatial and texture information from dual-temporal remote sensing pictures, we employ Siamese structures with weight sharing. The auxiliary modules include the Difference Feature Module (DFM), Attention Refinement Module (ARM), and Cross-Scale Fusion Module (CSFM). The decoder network is comprised of two parallel semantic branches. At each layer, the feature maps are processed in parallel by two consecutive convolution modules and two stripe convolution modules. The multi-dimensional semantic features and edge detail characteristics of the dual-temporal remote sensing pictures are then recovered by adding fusing to these processed feature maps. MDANet takes dual-temporal remote sensing images as input. These images are initially fed into the encoder network, where the backbone network extracts the basic features of the images. Subsequently, these features undergo further processing through DFM, ARM, and CSFM to enhance their representational capability. After processing by the encoder network, the feature maps are transmitted to the decoder network. In the decoder network, the feature maps pass through parallel semantic branches, undergo a series of convolutional operations, and are fused to produce the change detection results. During training, we employ supervised learning, updating the model parameters by comparing the differences between the model output and the ground truth labels to gradually optimize the model. Figure 1 illustrates the model's framework, where $E_{(i,j)}$ represents the encoder network, $i \in (1, 2)$ denotes the two branches of the decoder, $j \in (1, 2, 3, 4, 5)$ represents each encoder, and D_j represents the decoder network.

Encoder network: We alter the ResNet34 network during the feature encoding phase to make it more suited for our change detection objective. We use a dual-path approach for change feature extraction, with weight sharing between the two paths. Additionally, we eliminate the last two completely linked layers of ResNet34 as well as the average pooling layer. We utilize DFM to extract difference features at different levels between each encoder of ResNet34. In order to improve feature granularity and enhance the change features of dual-temporal remote sensing pictures, we utilize ARM to optimize the output features of each DFM. Because of this, the network is able to shift its channel and spatial emphasis. Considering that a simple and coarse fusion of features from different scales can impact

detection accuracy, following ARM, we use CFSM to capture context dependencies and integrate richer information. This makes it possible to concurrently encode high-level semantic information and low-level features, creating a more potent representation.

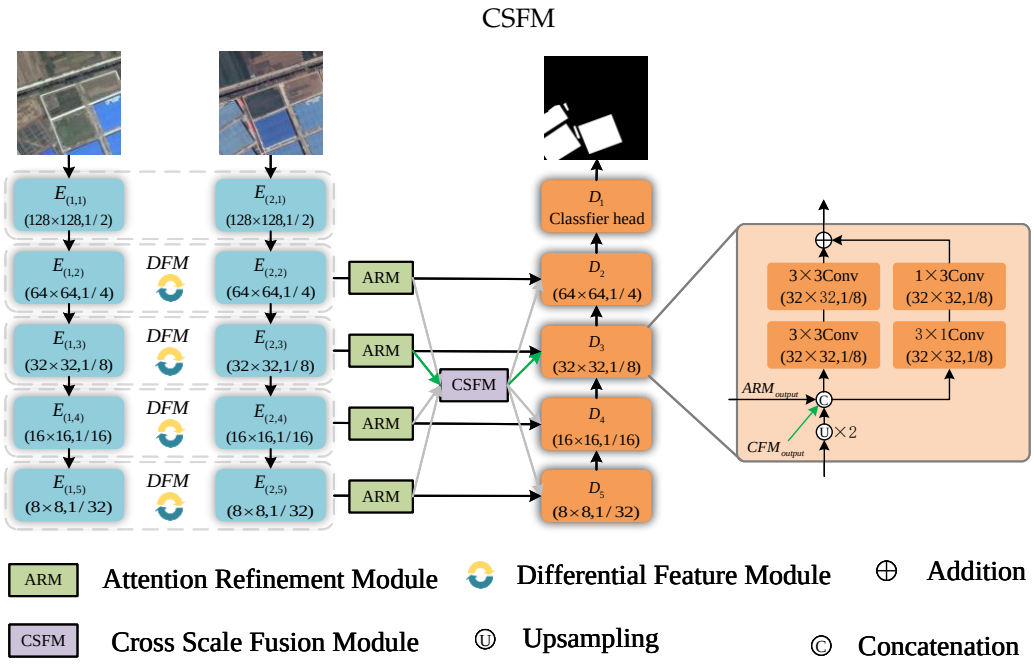

Figure 1. The overall framework of MDANet. Incorporating ResNet34 as the primary backbone network for decoding, the network has an encoder–decoder architecture.

Decoder network: The decoder network has five levels, much like the encoder network does. To guarantee that the output size of encoding blocks and the input size of decoding blocks at the same layer are the same, a two-fold downsampling frequency is used between encoder blocks, and a two-fold upsampling frequency is used between decoder blocks. Each decoder block has two parallel branches. One branch employs two 3×3 convolution layers to restore the image, while the other branch uses 3×1 and 1×3 stripe convolutions to extract edge detail information of the change regions. Square convolution windows extract too much irrelevant information, which can interfere with the model's predictive performance. Stripe convolution, as proposed by Hou [51], can reduce this interference from irrelevant factors. Stripe convolution is used to restore multi-scale spatial information in change areas, which enhances the detection of small change targets and the recovery of change boundaries.

3.1.2. Difference Feature Module (DFM)

The Difference Feature Module (DFM) is a new module that we have proposed in our research in the field of change detection. Its structural diagram is shown in Figure 2. In traditional change detection methods, subtle differences in images are often overlooked, leading to issues like missed detections and false alarms. Convolutional neural networks extract high-level features from dual-temporal remote sensing images, such as color, texture, and shape, which effectively characterize each region. Subsequently, through the Difference Feature Module, we compare the features from different time periods to measure the differences between them, thus determining which areas have changed. The Difference Feature Module is capable of effectively capturing minor changes, making the network more sensitive. This helps to lessen problems like edge blurring, false positives, and false negatives while also increasing the accuracy of change detection.

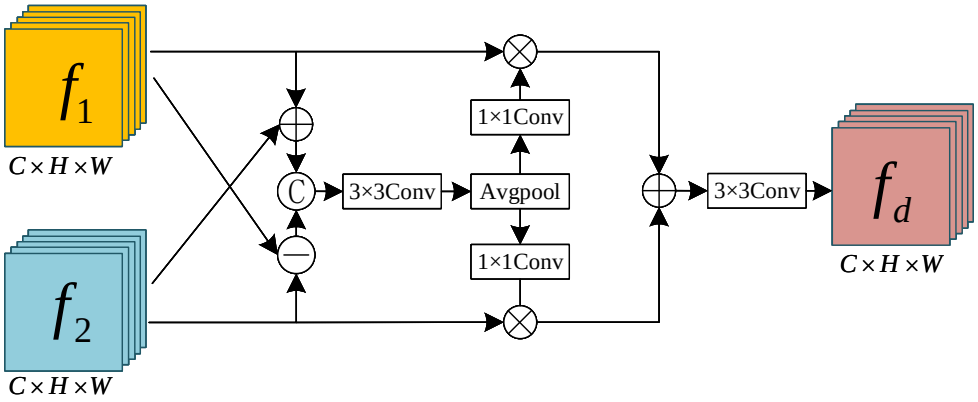

Figure 2. Difference Feature Module.

Assuming that f_1 and f_2 are two input features of the Difference Feature Module, firstly, we perform addition and absolute difference operations on f_1 and f_2 separately to obtain similar features and different features. Then, we concatenate the similar features and different features along the channel dimension and input them into a convolution operation on 3×3. After that, we apply average pooling to obtain the difference weight map f_w. Next, we feed f_w into a convolution operation on 1×1 and perform weighted fusion with f_1 and f_2 separately. Finally, the results are summed and input into a convolution operation on 3×3 to obtain the difference feature map f_d of the dual-temporal remote sensing image. The proposed Difference Feature Module (DFM) makes full use of the semantic and spatial information extracted by the main network at each layer, effectively distinguishing between changing and non-changing information, thus avoiding the issue of blurry detection boundaries and enhancing the efficiency of discriminative features. The following is a description of the computation formulae used in the method mentioned above:

$$f_w = f^{3\times3}[(f_1 + f_2); abs(f_1 - f_2)] \tag{1}$$

$$f_d = f^{3\times3}[f_1 \otimes f^{1\times1}(Avgpool(f_w)) + f_2 \otimes f^{1\times1}(Avgpool(f_w))] \tag{2}$$

In this context, $f^{n\times n}(\cdot)$ represents a convolutional layer with a kernel size of n, followed by batch normalization and the ReLU activation function [52]. $abs(\cdot)$ indicates performing the absolute difference operation. $[;]$ represents a concatenation operation along the channel dimension. $\otimes$ denotes element-wise multiplication. $Avgpool(\cdot)$ signifies average pooling.

3.1.3. Attention Refinement Module (ARM)

Due to the fact that dual-temporal remote sensing photos are taken at several times, they show variations in the seasons, illumination, and solar angles, which have a substantial impact on detection accuracy. Without effective focus on the changing regions, it becomes challenging for the network to distinguish each pixel and assign accurate labels. In this section, we propose a new Attention Refinement Module (ARM) as illustrated in Figure 3. This module aligns the spatial and channel dimensions for both changing and non-changing pixels, allowing the network to assign higher weights to changing regions and lower weights to non-changing areas, suppressing non-changing features and background regions. The Attention Refinement Module (ARM) accelerates and improves the training process by improving the network's capacity to acquire specifics and local information in the features. It also reduces issues like misclassification, omission, and edge blurring in changing regions.

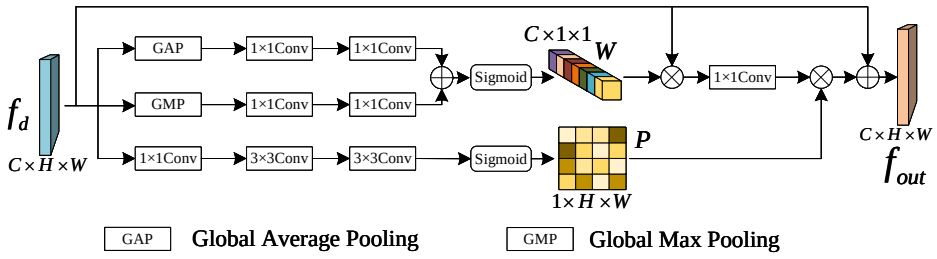

Figure 3. Attention Refinement Module.

The Attention Refinement Module (ARM) takes the output f_d from the Differential Feature Module (DFM) as input, with dimensions $C \times H \times W$ representing the channel dimension, height, and width of the feature maps, respectively. In order to create feature maps with dimensions $C \times 1 \times 1$, f_d is first processed by global average pooling and global max pooling, lowering the spatial dimension and improving the channel dimension data. Then, they go through two layers of 1×1 convolution and are added together to form a fused feature. The channel-wise weight coefficient matrix W is obtained using the sigmoid activation function. Next, we use convolution with 1×1 to fuse the input features f_d, compressing them into the spatial dimension, resulting in feature maps of size $1 \times H \times W$. These maps are further processed by two layers of 3×3 convolution to extract spatial feature information thoroughly. The spatial-wise weight coefficient matrix P is obtained using the sigmoid activation function. To prevent the loss of critical features during the network learning process, we introduce residual connections. We weight the fusion of the input feature f_d using the channel-wise weight coefficient matrix W. Afterward, the result goes through a single layer of convolution 1×1 and is then weighted again by the spatial-wise weight coefficient matrix P. Finally, it is added to the input feature f_d, resulting in the final output of the ARM, denoted as f_{out}. The mathematical computations for the described process are expressed as follows:

$$W = \sigma(f^{1\times1}(f^{1\times1}(GAP(f_d))) + f^{1\times1}(f^{1\times1}(GMP(f_d)))) \tag{3}$$

$$P = \sigma(f^{3\times3}(f^{3\times3}(f^{1\times1}(f_d)))) \tag{4}$$

$$f_{out} = f^{1\times1}(f_d \otimes W) \otimes P + f_d \tag{5}$$

In the above formula, $\sigma(\cdot)$ represents the sigmoid activation function, $GAP(\cdot)$ represents global average pooling, and $GMP(\cdot)$ represents global max pooling.

3.1.4. Cross-Scale Fusion Module (CSFM)

Features extracted through CNN shallow networks have higher resolution and are more visually rich compared to deep-layer features. Deep networks extract lower-resolution features that are more abstract and contain richer semantic information compared to shallow networks. However, as the network depth increases, information about small objects can be easily lost. This is because the inherent receptive fields of convolutional operations cannot effectively integrate contextual information. Although pooling operations partially improve this, the features lose resolution and some semantic information is lost as well. Shallow features in the context of high-resolution remote sensing image change detection typically include basic semantic information, such as the size and shape of the changed target areas, while deep features include more sophisticated semantic information, such as the distribution and spacing between various changed targets. Hence, fusing shallow and deep semantic features is crucial for change detection tasks. We propose a Cross-Scale Fusion Module (CSFM) as shown in Figure 4, which fuses features from different network layers. This enables the network to possess both shallow information about shapes and sizes, as well as deep semantic information. This increases the robustness of the model and strengthens deep learning networks' capacity for generalization.

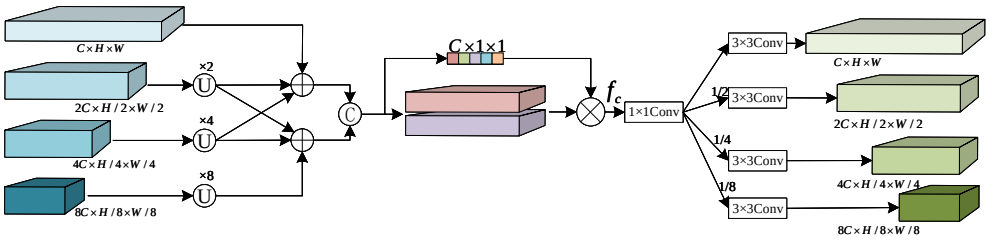

Figure 4. Cross-Scale Fusion Module.

The input to the Cross-Scale Fusion Module (CSFM) is represented as f_i, and it consists of the change information refined through ARM, where $i \in (1, 2, 3, 4)$. As the four input features of this module have different sizes and channel numbers, we upsample them to ensure they have the same size and channel numbers, resulting in the feature map $f_i^{'}$. However, because the relationships between pixels are diluted when the features from the lowest layer are directly propagated to the upper layer, and long-distance pixels lack correlation, we fuse the features from adjacent three layers to obtain $f_1^{''}$ and $f_2^{''}$. Then, we concatenate these two features in the channel dimension. We offer a channel attention technique to further the features' refinement and acquire f_c in order to eliminate redundant information from the learned features. We employ convolution on 1×1 for feature learning, followed by upsampling to restore the feature size to match the dimensions of the input features. The output f_{ci} is then produced by using another layer of 3×3 convolution to extract differential features and restore the channel numbers:

$$f_1^{''} = f_1 + Up^{\times 2}(f_2) + Up^{\times 4}(f_3) \tag{6}$$

$$f_2^{''} = Up^{\times 2}(f_2) + Up^{\times 4}(f_3) + Up^{\times 8}(f_4) \tag{7}$$

$$f_c = [f_1^{''}; f_2^{''}] \otimes \sigma(f^{1\times 1}(f^{1\times 1}(GAP([f_1^{''}; f_2^{''}])))) \tag{8}$$

$$f_{ci} = f^{3\times 3}(D^{\times j}(f^{1\times 1}(f_c))); i \in (1, 2, 3, 4); j \in (0, 2, 4, 8) \tag{9}$$

In the equation, $Up^{\times n}(\cdot)$ represents upsampling, $D^{\times n}(\cdot)$ represents downsampling, and n represents the scaling factor or ratio.

3.2. Datasets

We conducted a performance study on the three datasets BTCDD [53], LEVIR-CD [36], and CDD [54] to confirm the efficacy of our MDANet.

3.2.1. BTCDD

The BTCDD dataset is created from remote sensing images collected from Google Earth, spanning from 2010 to 2019. It includes a range of change objectives, including, among others, farms, highways, and buildings. The dataset consists of 3420 pairs of 512×512 pixel co-temporal remote sensing picture pairs in total. The dataset is divided into training, validation, and testing sets, containing 2280, 570, and 570 image pairs, respectively. In Figure 5, some samples from the dataset and their corresponding labels are displayed, illustrating that the dataset encompasses a wide range of common application scenarios.

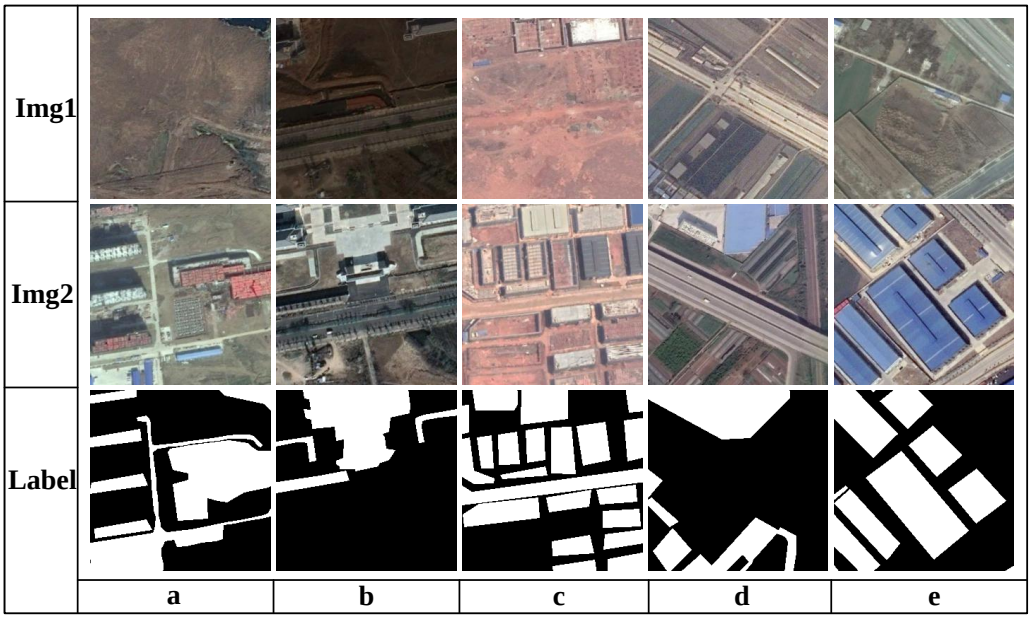

Figure 5. Graph representing the BTCDD dataset. Co-temporal remote sensing pictures are shown in the first and second rows, and their associated labels are shown in the third row.

3.2.2. LEVIR-CD

The Beihang University LEVR Laboratory has made available the LEVIR-CD dataset, which is a sizable remote sensing building change detection dataset. This collection consists of 637 pairs of 1024 × 1024 pixel ultra-high resolution photos. Domain experts in remote sensing image interpretation annotated this dataset, categorizing it into two classes: change and unchanged. Figure 6 illustrates sample images from the LEVIR-CD dataset. The images in the dataset are resized to 256 × 256 pixels, and there are a total of 10,192 image pairs. The training set includes 7120 image pairs, the validation set includes 1024 image pairs, and the test set includes 2048 image pairs.

Figure 6. Graph representing the LEVIR-CD dataset. Co-temporal remote sensing pictures are shown in the first and second rows, and their associated labels are shown in the third row.

3.2.3. CDD

The 11 pairs of remote sensing photos with seasonal changes that make up CDD are a public dual-temporal remote sensing image change detection data collection. Seven of the pairings have dimensions of 4725 × 2700, while the other four have dimensions of 1900 × 1000. The collection contains seasonal variations in the vegetation along with features, buildings, and forest regions of various sizes. Concurrently, the dataset is divided into 256 × 256 picture blocks, and a fresh change detection dataset comprising 16,000 dual-phase image pairs is created. There are 10,000 pairs in the training set, and 3000 pairs in each of the test and verification sets. A schematic design of a few examples from the CDD dataset may be seen in Figure 7.

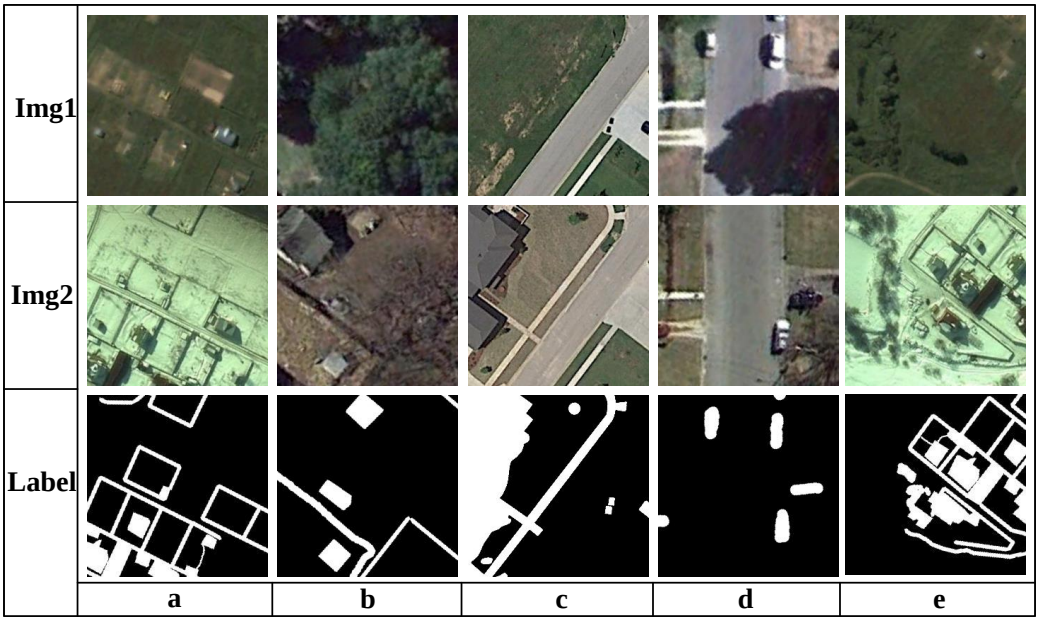

Figure 7. Graph representing the CCD dataset. Co-temporal remote sensing pictures are shown in the first and second rows, and their associated labels are shown in the third row.

3.3. Implementation Details

3.3.1. Evaluation Metrics

This study presents a number of objective quantitative criteria for the objective evaluation of the performance of various change detection techniques. By comparing the predictions made by various algorithms, these measures analytically evaluate algorithm performance differences. In this study, we employ precision (PR), recall (RC), intersection over union (IoU), and F1 score ($F1$) to evaluate various models. $F1$ and IoU are used as the primary evaluation metrics. $F1$ is the weighted harmonic mean of precision (PR) and recall (RC), providing a comprehensive assessment of both metrics. IoU measures the ratio of the intersection of predicted results for a specific class and the ground truth to their union:

$$PR = \frac{TP}{TP + FP} \tag{10}$$

$$RC = \frac{TP}{TP + FN} \tag{11}$$

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

$$F1 = \frac{2 \times PR \times RC}{PR + RC} \tag{13}$$

Among these, TP denotes the proportion of positive samples that were properly predicted, FP the proportion of positive samples that were wrongly predicted, and FN the proportion of negative samples that were mistakenly predicted.

3.3.2. Experimental Details

Based on the Pytorch framework, the GeForce RTX 3080 is used for training and testing in all of the experiments in this paper. The storage size of the GPU is 10 GB. We employ the Ploy technique to dynamically modify the learning rate because learning rate is essential to the entire training procedure. The formula for its mathematical computation is as follows:

$$lr = lr^* \times (1 - \frac{epoch}{n\_epoch})^p \tag{14}$$

Among them, lr represents the current learning rate, lr^* represents the initial learning rate, $epoch$ represents the current number of iterations, n_epoch represents the total number of iterations, and p represents the constant that controls the decay rate. In this article, we set the training batch size to 8, lr^* to 0.0015, n_epoch to 200, and p to 0.9. The loss function is BCEWithLogitsLoss, and Adam is the optimization method.

4. Results

On two datasets, we perform generalization, comparison, and ablation experiments in this section.

4.1. Network Structure Selection

ResNet is selected as our backbone network because of its deep network structure, which enables the network to learn more complicated features and representations. This is very useful for tackling picture change detection issues. Deep networks can capture more abstract and high-level features, contributing to improved algorithm performance. Additionally, ResNet introduces residual connections that enable smoother information flow throughout the network. This helps mitigate the problem of vanishing gradients, enhances training stability, and accelerates convergence. In change detection, this is crucial for accurately capturing change information. On the BTCDD dataset, we evaluate the performance of ResNet18, ResNet34, and ResNet50. ResNet34 displays the best performance as shown in Table 1. As a result, ResNet34 serves as the backbone network.

Table 1. Comparison results of ResNet at different depths (bold indicates the best).

Method	PR (%)	RC (%)	IoU (%)	F1 (%)
ResNet18	88.63	77.15	70.19	82.46
ResNet34	**89.57**	**79.52**	**72.79**	**84.25**
ResNet50	87.95	78.11	71.07	82.70
ResNet101	87.66	78.98	71.09	83.12

4.2. Ablation Experiments

We incrementally connect the DFM, ARM, and CSFM modules to the backbone network and run ablation tests on the BTCDD dataset to confirm the efficacy of the three modules we suggest. All experiments adopt the same strategy. The experimental results are as follows as shown in Table 2.

(1) Ablation experiments for DFM: DFM addresses the issues of omission or misjudgment caused by direct feature subtraction in traditional change detection. This module effectively captures subtle changes, reducing interference from lighting, seasonal, and angle variations, thus making the network more sensitive and significantly enhancing change detection accuracy. As shown in Table 2, the experimental results indicate that by adding DFM to the backbone network, IoU and F1 scores improve by 1.6% and 1.09%, respectively.

(2) Ablation experiments for ARM: ARM enables the network to focus more intensively on processing change regions by increasing the weight of change areas while reducing the weight of redundant information. It devotes more time to change characteristics, improving how change information is perceived and handled. The experimental results in Table 2 show that by adding ARM to the backbone network, IoU and F1 scores improve by 1.57% and 1.07%, respectively.

(3) Ablation experiments for CSFM: Basic semantic information, like the size and shape of shallow change areas, and more complex semantic information, such the distribution and spacing of various change targets in deep layers, are concurrently encoded by CSFM. This reduces the model's dependency on specific information, strengthens its ability to extrapolate, and enhances how it perceives and uses traits from various scales. The results in Table 2 indicate that by adding CSFM to the backbone network, IoU and F1 scores increase by 0.61% and 0.41%, respectively.

The results of the ablation tests performed on our proposed DFM, ARM, and CSFM modules demonstrate that, for change detection tasks, gradually integrating these modules lets the network effectively extract change characteristics from remote sensing images. Figure 8 shows the prediction effect of adding DFM, ARM and CSFM modules to the three datasets in sequence. It is obvious that these three modules gradually improve the performance of the model in extracting edge detail information and improve the detection ability of small targets.

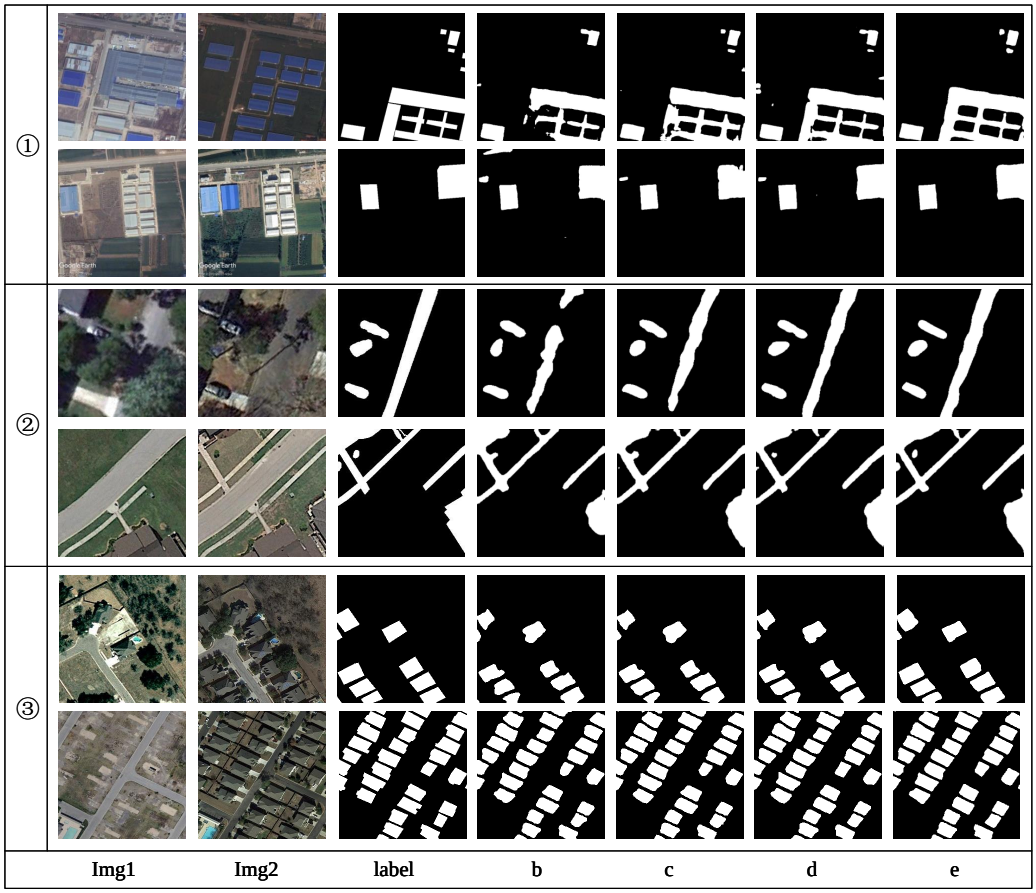

Figure 8. Visualization renderings of ablation experiments. We conduct experiments on three datasets: BTCDD, CDD and LEVIR-CD. b is the prediction effect of Backbone; c is the prediction effect of Backbone + DFM; d is the prediction effect of Backbone + DFM + ARM; e is the prediction effect of Backbone + DFM + ARM + CSFM.

Table 2. Ablation comparison experimental results of MDANet (bold indicates the best).

Method	PR (%)	RC (%)	IoU (%)	F1 (%)
Backbone	86.41	77.41	69.01	81.68
Backbone+DFM	88.76	77.54	70.61	82.77
Backbone+DFM+ARM	89.21	79.08	72.18	83.84
Backbone+DFM+ARM+CSFM	**89.57**	**79.52**	**72.79**	**84.25**

4.3. Comparative Experiments of Different Algorithms on BTCDD

In this part, we compare our proposed strategy against a number of change detection methods to show its superiority. The experimental results are shown in Table 3. To assure the experiment's objectivity and impartiality, we apply the same methodology to all of the models. The findings in the table show that, in addition to the approach we suggest, ChangeFormer has the best experimental outcomes, with IoU and F1 scores reaching 71.29% and 83.24% respectively, and our algorithm is the best in all four indicators. The IoU and F1 scores reach 72.79% and 84.25%, indicating that our MDANet method has certain advantages compared with other methods.

Table 3. Comparative test on BTCDD (bold indicates the best).

Method	PR (%)	RC (%)	IoU (%)	F1 (%)	FLOPs (G)	Time (ms)
FC-EF [46]	77.46	43.75	38.81	55.91	**5.43**	7.36
FC-Siam-Diff [46]	77.63	46.26	40.82	57.97	8.97	**5.06**
FC-Siam-Conc [46]	82.78	43.28	42.21	59.37	9.48	5.32
TCDNet [55]	88.64	74.37	67.91	80.89	7.96	8.51
SNUNet [56]	84.68	78.82	68.98	81.64	96.67	8.73
STANet [36]	86.55	77.36	69.06	81.69	53.03	19.45
DASNet [37]	87.59	77.22	69.61	82.08	103.54	17.62
ChangNet [57]	88.27	76.81	69.69	82.15	40.36	15.98
TFI-GR [58]	88.98	76.38	69.78	82.19	17.78	12.53
BIT [38]	87.23	78.66	70.52	82.73	98.61	8.96
MFGAN [59]	88.65	77.97	70.89	82.97	49.74	8.95
ChangeFormer [60]	87.81	79.13	71.29	83.24	78.47	9.62
MDANet (our)	**89.57**	**79.52**	**72.79**	**84.25**	7.42	15.21

The aforesaid method's prediction graph for the BTCDD dataset is shown in Figure 9. Three sets of dual-temporal remote sensing images from distinct locations are compared in order to further support the practicality of our proposed approach. In the figure, 'a' represents the image label, and 'b–l' represent the prediction map of the above method. As can be observed in the figure, the prediction maps of other algorithms have serious problems with false detection, missing detection, and the blurring of changing objects. In contrast, our method effectively avoids these problems and improves edge details. It processes tiny target shifting regions extremely well, and the prediction result map is most similar to the picture label, which attests to the potency of our suggested approach.

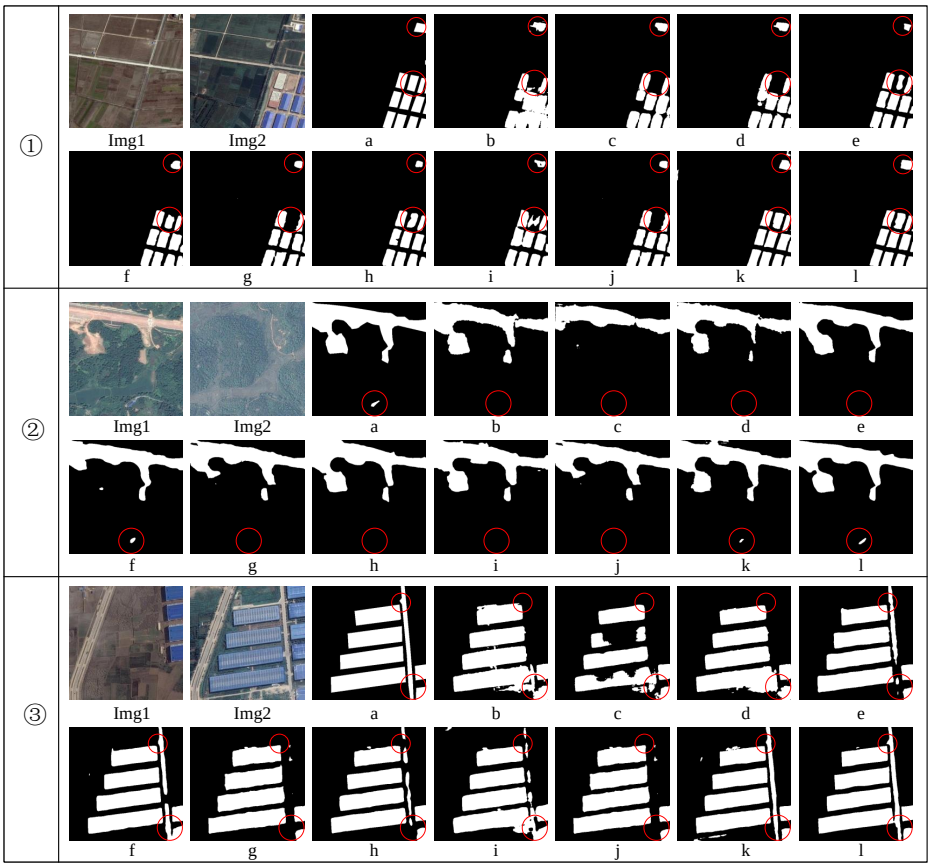

Figure 9. Prediction effects diagram of different algorithms on BTCDD. We conduct experiments using three pairs of dual-temporal remote sensing images. Img1 and Img2 represent sets of sense images in different periods, (a–l) respectively represent label, FC-EF, FC-Siam-Diff, FC-Siam-Conc, TCDNet, DASNet, ChangNet, TFI-GR, BIT, MFGAN, ChangaFormer, MIDANet (Ours).

4.4. Generalization Experiments of Different Algorithms on LEVIR-CD

We conduct generalization tests on the LEVIR-CD dataset to illustrate the generalization performance of our suggested technique, and the results are shown in Table 4. Among these methods, aside from our proposed method, ChangeFormer achieves the best results with IoU and F1 scores of 81.24% and 89.65%, respectively. However, our MDANet method achieves IoU and F1 scores of 82.94% and 90.67% on this dataset, surpassing ChangeFormer by 1.7% and 1.02%, highlighting the strong robustness and generalization capabilities of our approach.

Table 4. Generalization experiments in LEVIR-CD. (The best results are indicated in bold font).

Method	PR (%)	RC (%)	IoU (%)	F1 (%)
FC-EF	84.12	83.82	72.37	83.97
FC-Siam-Diff	87.69	81.74	73.33	84.61
FC-Siam-Conc	88.27	84.06	75.61	86.11
TCDNet	89.99	85.27	77.88	87.57
SNUNet	90.67	85.52	78.61	88.02
STANet	90.82	85.97	78.95	88.23
DASNet	89.93	82.92	78.38	87.88
ChangNet	87.81	89.12	79.31	88.46
TFI-GR	89.59	89.15	80.78	89.37
BIT	90.94	87.29	80.31	89.08
MFGAN	88.98	89.01	80.17	88.99
ChangeFormer	89.57	89.73	81.24	89.65
MDANet (our)	**90.99**	**90.35**	**82.94**	**90.67**

The prediction outcomes of the aforementioned change detection techniques on the LEVIR-CD dataset are shown in Figure 10. To further confirm the effectiveness of our suggested method's generalization, we compare it with these approaches using three pairs of co-temporal remote sensing picture pairings from various places. In the figure, 'a' represents the image label, while 'b–l' represent the prediction images of the above-mentioned methods. The images show that our suggested MDANet technique offers more precise predictions for tiny targets and edge details. The predicted change regions are also clearer, even for very close change areas, demonstrating the ability to distinctly identify adjacent change regions with clear boundaries. This illustrates the superiority of our MDANet approach.

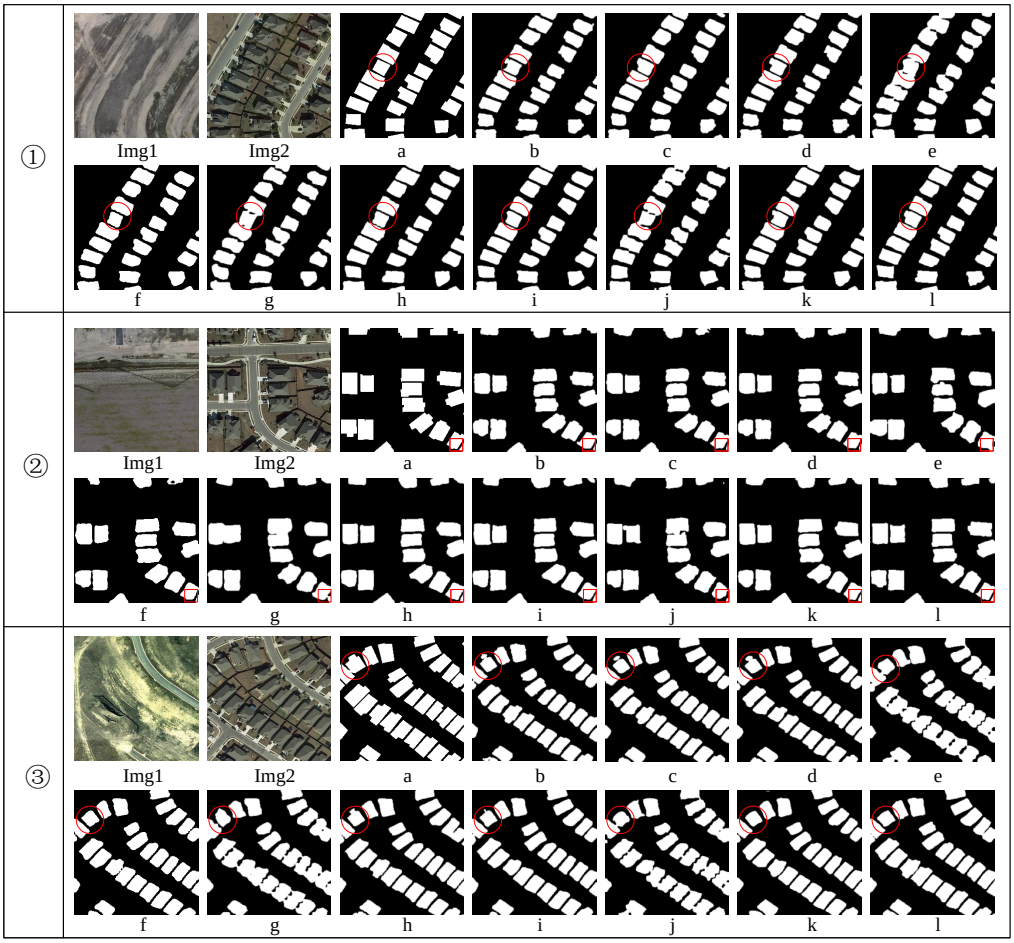

Figure 10. Prediction effect diagram of different algorithms on LEVIR-CD. We conduct experiments using three pairs of dual-temporal remote sensing images. Img1 and Img2 represent sets of sense images in different periods, (a–l) respectively represent label, FC-EF, FC-Siam-Diff, FC-Siam-Conc, TCDNet, DASNet, ChangNet, TFI-GR, BIT, MFGAN, ChangaFormer, MIDANet (Ours).

4.5. Generalization Experiments of Different Algorithms on CDD

Using the CDD dataset, we further confirm the MDANet generalization performance. The experimental outcomes are displayed in Table 5. The table shows that overall, our suggested approach produces the greatest outcomes, with IoU and F1 scores of 81.18% and 89.61%, respectively. Additionally, besides our method, ChangeFormer achieves IoU and F1 scores of 80.25% and 89.04% on this dataset, respectively. Our MDANet method achieves IoU and F1 scores 0.93% and 0.57% higher than ChangeFormer on this dataset, demonstrating the superior generalization performance of our proposed approach.

Table 5. Generalization experiments in CDD. (The best results are indicated in bold font).

Method	PR (%)	RC (%)	IoU (%)	F1 (%)
FC-EF	79.79	61.28	53.05	69.32
FC-Siam-Diff	74.83	70.64	57.08	72.67
FC-Siam-Conc	79.49	65.05	55.70	71.55
TCDNet	84.39	89.79	77.01	87.01
SNUNet	84.85	89.82	77.41	87.26
STANet	83.32	91.35	78.08	87.69
DASNet	83.49	91.12	77.21	87.14
ChangNet	83.34	89.31	75.78	86.22
TFI-GR	83.78	93.84	79.42	88.53
BIT	83.52	93.95	79.26	88.43
MFGAN	83.36	92.26	77.91	87.58
ChangeFormer	84.69	93.86	80.25	89.04
MDANet(our)	**85.59**	**94.03**	**81.18**	**89.61**

The aforesaid change detection method's prediction effect on the CDD dataset is depicted in Figure 11. To further confirm the generalization performance of our suggested strategy, we compare it on three pairs of dual-temporal remote sensing photos from various locations. The prediction map of the aforementioned approach is shown by $b--l$ in the picture, while a stands for the image label. The graphic illustrates how our approach performs the best, as it takes advantage of the attention mechanism to improve the correlation between channels and pixels, which improves the method's ability to extract altered regions and distinguish boundaries. Additionally, it lowers missed detections. The best detection results are obtained, demonstrating our suggested method's superior generalization capability.

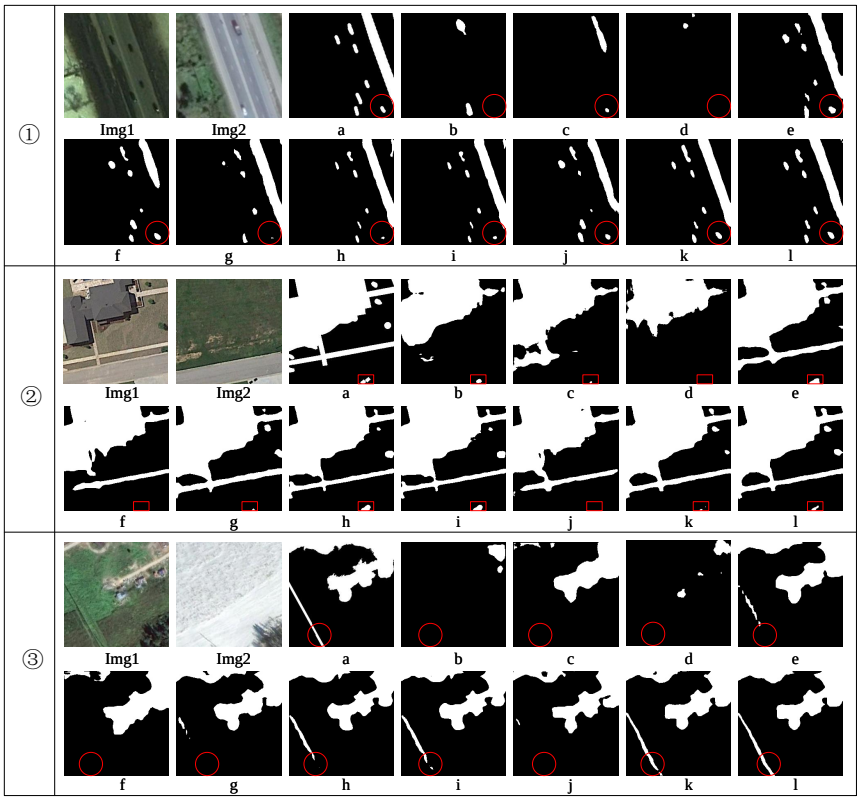

Figure 11. Prediction effect diagram of different algorithms on CDD. We conduct experiments using three pairs of dual-temporal remote sensing images. Img1 and Img2 represent sets of sense images in different periods, (a–l) respectively represent label, FC-EF, FC-Siam-Diff, FC-Siam-Conc, TCDNet, DASNet, ChangNet, TFI-GR, BIT, MFGAN, ChangaFormer, MIDANet (Ours).

5. Discussion

In comparative experiments and generalization experiments, our proposed method is better than other methods in terms of evaluation indicators and prediction renderings, and it can be seen from the prediction renderings that our method can effectively detect dual-phase geographically changing areas in remote sensing images. From the ablation experiments in Figure 8, it can be observed that with the sequential integration of DFM, ARM, and CSFM modules into the backbone network, the predicted result gradually exhibits a trend of increased accuracy and clarity. Firstly, after the incorporation of the DFM, the change regions in the predicted result exhibit more distinct features, enabling the model to better capture the differential information between images, thereby effectively enhancing the accuracy of change detection, as shown in Figure 8c. Secondly, with the addition of ARM, the edges of the change regions in the predicted result become clearer, and the identification of small targets and subtle changes is improved, further enhancing the model's learning capability of image details and local information. Figure 12 demonstrates the heat map effect of ARM, where (a) and (b) represent the heat maps produced when ARM is removed and added to the network, respectively. From column (a), it can be seen that the effect on edge details and small targets is poor when the network removes the ARM module, and there is a tendency to detect two adjacent change regions as one when they are very close. However, with the inclusion of ARM in the network, the detection performance is significantly improved, particularly for edge details and small targets. The attention mechanism of ARM has significant value in change detection tasks, as it dynamically adjusts the network's focus on features, thereby enhancing the model's perception of important information. Compared to Transformer, it can directly operate on feature maps without introducing additional attention matrix calculations as shown in Figure 8d. Lastly, the introduction of CSFM further improves the edge details of change regions in the predicted result, making the boundaries of change regions clearer and more precise. By fully utilizing the feature information between different levels, the model can better restore the edge details of change regions, thereby improving the accuracy and robustness of change detection as shown in Figure 8e. Experimental results on three datasets, BTCDD, LEVIR-CD and CDD, prove the effectiveness and superiority of our method. In the BTCDD dataset, our MDANet method achieves 82.94% and 90.67% IoU and F1 scores on this dataset. The prediction rendering is also the closest to the label, while other methods such as FC-Siam-Diff have serious changes in targets. In contrast, our method effectively mitigates problems such as false detection, missed detection, and blurring. The generalization experiment on the LEVIR-CD dataset also proves the generalizability of our method. On this dataset, the evaluation index IoU and F1 scores reach 82.94% and 90.67%. It can be seen from the prediction renderings that the overall prediction effects of different models are better, but in terms of the prediction of edge details and small targets, our method is more accurate and can successfully segment adjacent changing areas with clear boundaries. For example, MFGAN cannot predict two very close change areas, and there is a false detection phenomenon. The generalization performance of our MDANet is further verified on the CDD dataset. The IoU and F1 scores reach 81.18% and 89.61%, respectively, and it can be clearly seen from the prediction effect image that our method has the best effect because we use the attention mechanism. Strengthening the correlation between channels and pixels makes our method better at extracting changed areas and clearer boundary identification, while methods such as TFI-GR and BIT have obvious missed detections. Furthermore, our model demonstrates good efficiency in handling complex tasks, capable of rapidly and accurately detecting changes in the Earth's surface, while also performing well under resource-constrained conditions. The relatively small size of the model enables it to be more flexible and convenient to deploy and use.

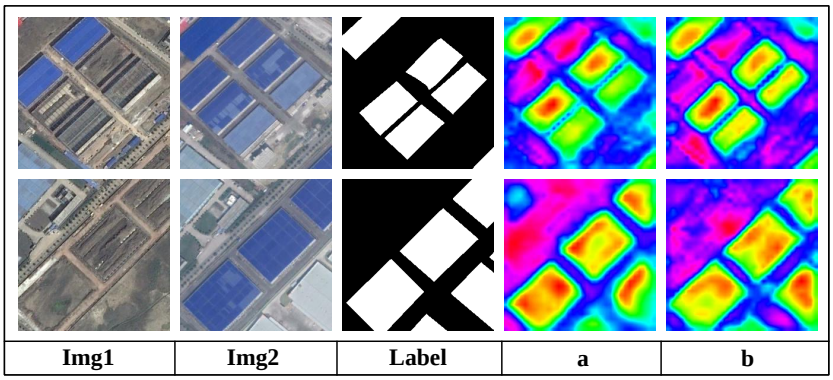

Figure 12. Comparison of ARM Heatmaps. The first and second columns each show a pair of co-temporal remote sensing pictures, while the third column displays the labels for each image pair. Heatmap (a) represents the heatmap generated when the network excludes the ARM, and heatmap (b) represents the heatmap generated when the network includes the ARM. The warm-colored regions represent the areas where the network focuses its attention, while the cold-colored regions indicate the areas suppressed by the network.

6. Conclusions

In this paper, we propose a high-resolution city change detection network based on difference and attention mechanisms under multi-scale feature fusion (MDANet). We have constructed this network using an encoder–decoder architecture. Additionally, we have proposed three auxiliary modules: DFM, ARM, and CSFM. DFM measures the differences between features from different temporal instances by comparing them, determining which regions have undergone changes. ARM allocates larger weights to change regions and smaller weights to non-change regions, suppressing non-change features and background areas. This allows the network to comprehensively learn details and local information from the features. By combining characteristics from several levels, CSFM enables the network to take into account both superficial information, such shape and size, and deep semantic information. This improves the deep learning network's capacity to generalize the model and robustness. By incrementally integrating our proposed modules into the backbone network, we observe a significant improvement in both IoU and F1 scores, indicating the effectiveness of our modules. Results from experiments show that MDANet performs better than other algorithms on the BTCDD, LEVIR-CD and CDD datasets. In the context of urban environment classification, our algorithm has demonstrated excellent performance, effectively identifying and classifying various features and targets within urban areas, including buildings, roads, green spaces, and detecting changes in these features. We believe that with further research and validation, our algorithm may also be applicable to other environmental classification tasks, such as rural areas, forest regions, and beyond. At the same time, in future change detection research, we will pay more attention to the fusion of multi-source data, including satellites, drones, sensors, etc, to improve accuracy and reliability. Moreover, most current change detection methods focus on the research of binary change detection. Although such methods can automatically monitor and analyze the changed areas in multi-temporal remote sensing images, they cannot describe the specific change types of ground objects. Therefore, in the future, how to interpret the specific types of ground objects before and after changes from remote sensing images will need to be further explored. Generally speaking, future research in the field of change detection will focus on improving algorithm accuracy, applying multi-source data fusion and deep learning technology, improving real-time performance, and refined monitoring to meet the needs of different fields.

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

Funding: This work is supported in part by the National Natural Science Foundation of PR China (42075310).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Bala, G.; Caldeira, K.; Wickett, M.; Phillips, T.; Lobell, D.; Delire, C.; Mirin, A. Combined climate and carbon-cycle effects of large-scale deforestation. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 6550–6555. [CrossRef] [PubMed]
2. Wang, W.; Liu, H.; Li, Y.; Su, J. Development and management of land reclamation in China. *Ocean Coast. Manag.* **2014**, *102*, 415–425. [CrossRef]
3. Trenberth, K.E. Climate change caused by human activities is happening and it already has major consequences. *J. Energy Nat. Resour. Law* **2018**, *36*, 463–481. [CrossRef]
4. Bruzzone, L.; Bovolo, F. A novel framework for the design of change-detection systems for very-high-resolution remote sensing images. *Proc. IEEE* **2012**, *101*, 609–630. [CrossRef]
5. Li, D.; Yan, S.; Zhao, M.; Chow, T.W. Spatiotemporal tree filtering for enhancing image change detection. *IEEE Trans. Image Process.* **2020**, *29*, 8805–8820. [CrossRef] [PubMed]
6. Radke, R.J.; Andra, S.; Al-Kofahi, O.; Roysam, B. Image change detection algorithms: A systematic survey. *IEEE Trans. Image Process.* **2005**, *14*, 294–307. [CrossRef]
7. 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]
8. 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]
9. Schmitt, A.; Wessel, B.; Roth, A. Curvelet-based change detection on SAR images for natural disaster mapping. *Photogramm. Fernerkund. Geoinf.* **2010**, *2010*, 463–474. [CrossRef]
10. 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]
11. D'Addabbo, A.; Refice, A.; Pasquariello, G.; Lovergine, F.P.; Capolongo, D.; Manfreda, S. A Bayesian network for flood detection combining SAR imagery and ancillary data. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3612–3625. [CrossRef]
12. Du, Y.; Teillet, P.M.; Cihlar, J. Radiometric normalization of multitemporal high-resolution satellite images with quality control for land cover change detection. *Remote Sens. Environ.* **2002**, *82*, 123–134. [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] [PubMed]
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. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
16. Meyer, D.; Wien, F. Support vector machines. *R News* **2001**, *1*, 23–26.
17. Volpi, M.; Tuia, D.; Bovolo, F.; Kanevski, M.; Bruzzone, L. Supervised change detection in VHR images using contextual information and support vector machines. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *20*, 77–85. [CrossRef]
18. Wang, X.; Liu, S.; Du, P.; Liang, H.; Xia, J.; Li, Y. Object-based change detection in urban areas from high spatial resolution images based on multiple features and ensemble learning. *Remote Sens.* **2018**, *10*, 276. [CrossRef]
19. Gong, M.; Cao, Y.; Wu, Q. A neighborhood-based ratio approach for change detection in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2011**, *9*, 307–311. [CrossRef]
20. Liu, H.; Yang, M.; Chen, J.; Hou, J.; Deng, M. Line-constrained shape feature for building change detection in VHR remote sensing imagery. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 410. [CrossRef]
21. Desclée, B.; Bogaert, P.; Defourny, P. Forest change detection by statistical object-based method. *Remote Sens. Environ.* **2006**, *102*, 1–11. [CrossRef]
22. Liu, J.; Gong, M.; Qin, K.; Zhang, P. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *29*, 545–559. [CrossRef] [PubMed]
23. Weismiller, R.; Kristof, S.; Scholz, D.; Anuta, P.; Momin, S. Change detection in coastal zone environments. *Photogramm. Eng. Remote Sens.* **1977**, *43*, 1533–1539.

24. 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]
25. Bovolo, F.; Bruzzone, L. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Trans. Geosci. Remote Sens.* **2006**, *45*, 218–236. [CrossRef]
26. Chen, Q.; Chen, Y. Multi-feature object-based change detection using self-adaptive weight change vector analysis. *Remote Sens.* **2016**, *8*, 549. [CrossRef]
27. Luppino, L.T.; Bianchi, F.M.; Moser, G.; Anfinsen, S.N. Unsupervised image regression for heterogeneous change detection. *arXiv* **2019**, arXiv:1909.05948.
28. 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]
29. Wei, D.; Hou, D.; Zhou, X.; Chen, J. Change detection using a texture feature space outlier index from mono-temporal remote sensing images and vector data. *Remote Sens.* **2021**, *13*, 3857. [CrossRef]
30. Shafique, A.; Cao, G.; Khan, Z.; Asad, M.; Aslam, M. Deep learning-based change detection in remote sensing images: A review. *Remote Sens.* **2022**, *14*, 871. [CrossRef]
31. Ding, A.; Zhang, Q.; Zhou, X.; Dai, B. Automatic recognition of landslide based on CNN and texture change detection. In Proceedings of the 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, China, 11–13 November 2016; IEEE: Toulouse, France, 2016; pp. 444–448.
32. Hou, B.; Wang, Y.; Liu, Q. Change detection based on deep features and low rank. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2418–2422. [CrossRef]
33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
34. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 3–13. [CrossRef]
35. Zhang, W.; Lu, X. The spectral-spatial joint learning for change detection in multispectral imagery. *Remote Sens.* **2019**, *11*, 240. [CrossRef]
36. 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]
37. 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]
38. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
39. Chen, P.; Zhang, B.; Hong, D.; Chen, Z.; Yang, X.; Li, B. FCCDN: Feature constraint network for VHR image change detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 101–119. [CrossRef]
40. Shen, Q.; Huang, J.; Wang, M.; Tao, S.; Yang, R.; Zhang, X. Semantic feature-constrained multitask siamese network for building change detection in high-spatial-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *189*, 78–94. [CrossRef]
41. Liao, C.; Hu, H.; Yuan, X.; Li, H.; Liu, C.; Liu, C.; Fu, G.; Ding, Y.; Zhu, Q. BCE-Net: Reliable building footprints change extraction based on historical map and up-to-date images using contrastive learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *201*, 138–152. [CrossRef]
42. 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.
43. 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]
44. Alcantarilla, P.F.; Stent, S.; Ros, G.; Arroyo, R.; Gherardi, R. Street-view change detection with deconvolutional networks. *Auton. Robot.* **2018**, *42*, 1301–1322. [CrossRef]
45. Peng, D.; Zhang, Y.; Guan, H. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
46. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; IEEE: Toulouse, France, 2018; pp. 4063–4067.
47. Guo, E.; Fu, X.; Zhu, J.; Deng, M.; Liu, Y.; Zhu, Q.; Li, H. Learning to measure change: Fully convolutional siamese metric networks for scene change detection. *arXiv* **2018**, arXiv:1810.09111.
48. Hussain, M.; Chen, D.; Cheng, A.; Wei, H.; Stanley, D. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS J. Photogramm. Remote Sens.* **2013**, *80*, 91–106. [CrossRef]
49. Shi, W.; Zhang, M.; Zhang, R.; Chen, S.; Zhan, Z. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sens.* **2020**, *12*, 1688. [CrossRef]
50. 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.

51. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4003–4012.
52. Jiang, S.; Dong, R.; Wang, J.; Xia, M. Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network. *Systems* **2023**, *11*, 305. [CrossRef]
53. 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]
54. Lebedev, M.; Vizilter, Y.V.; Vygolov, O.; Knyaz, V.A.; Rubis, A.Y. Change detection in remote sensing images using conditional adversarial networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 565–571. [CrossRef]
55. 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]
56. 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]
57. 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.
58. 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]
59. Chu, S.; Li, P.; Xia, M. MFGAN: Multi feature guided aggregation network for remote sensing image. *Neural Comput. Appl.* **2022**, *34*, 10157–10173. [CrossRef]
60. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. In Proceedings of the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Toulouse, France, 2022; pp. 207–210.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.