remote sensing

MDPI

Technical Note

Change Detection for High-Resolution Remote Sensing Images Based on a Multi-Scale Attention Siamese Network

Jiankang Li [1], Shanyou Zhu [1,*], Yiyao Gao [1], Guixin Zhang [2] and Yongming Xu [1]

[1] School of Remote Sensing & Geomatics Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China; jkang_li@nuist.edu.cn (J.L.); gyy995522@nuist.edu.cn (Y.G.); xym30@nuist.edu.cn (Y.X.)
[2] School of Geographical Science, Nanjing University of Information Science & Technology, Nanjing 210044, China; 001631@nuist.edu.cn
* Correspondence: zsyzgx@nuist.edu.cn

Abstract: To address the problems in remote sensing image change detection such as missed detection of features at different scales and incomplete region detection, this paper proposes a high-resolution remote sensing image change detection model (Multi-scale Attention Siamese Network, MASNet) based on a Siamese network and multi-scale attention mechanism. The MASNet model took the Siamese structure of the ResNet-50 network to extract features of different simultaneous images and then applied the attention module to feature maps of different scales to generate multi-scale feature representations. Meanwhile, an improved contrastive loss function was adopted to enhance the learning of change features and improving the imbalance problem between unchanged and changed samples. Furthermore, to address the current time-consuming and laborious phenomenon of manually annotating datasets, we provided a change detection dataset from Yunnan Province in China (YNCD) that contains 1540 pairs of 256 × 256 bi-temporal images with a spatial resolution of 1 m. Then, model training and change detection applications were studied by expanding a small number of experimental area samples into the existing public datasets. The results showed that the overall accuracy of the MASNet model for change detection in the experimental area is 95.34%, precision rate is 79.78%, recall rate is 81.52%, and F1 score is 80.64%, which are better than those of six comparative models (FC-EF, FC-Siam-Diff, FC-Siam-Conc, PAN, MANet, and STANet). This verifies the effectiveness of the MASNet model as well as the feasibility of change detection by expanding existing public datasets.

Keywords: change detection; deep learning; Siamese network; attention mechanism; Yun-nan datasets

check for updates

Citation: Li, J.; Zhu, S.; Gao, Y.; Zhang, G.; Xu, Y. Change Detection for High-Resolution Remote Sensing Images Based on a Multi-Scale Attention Siamese Network. *Remote Sens.* **2022**, *14*, 3464. https://doi.org/10.3390/rs14143464

Academic Editors: Kamil Krasuski and Damian Wierzbicki

Received: 9 June 2022
Accepted: 18 July 2022
Published: 19 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Remote sensing image (RSI) change detection (CD) refers to the comparison and analysis of images of the same region in different periods through image processing and other means to judge the changes between images [1]. With the in-depth development of aerospace technology and electronic information technology, the resolution of remote sensing images is continuously improving. The demand for large-scale land cover change or specified element changes using optical remote sensing images with high spatial resolution are increasing.

According to different research objects, traditional CD methods can be divided into pixel-based and object-based change detection methods [2]. The pixel-based change detection method generates a difference image by directly comparing the spectral information or texture information of the pixels and obtains the final change result map through threshold segmentation or clustering [3–5]. Although this method is simple and feasible, it ignores the spatial background information and generates considerable "salt and pepper" noise during processing. The object-based change detection method divides the remote sensing image into disjoint objects and analyzes the differences between spatial-temporal images through

the rich spectral, texture, structure, and geometric information [6–9]. This method utilizes the spatial background information of high-resolution RSI, but the artificial feature extraction process is complicated and shows poor robustness. Wu et al. combined pixel-level and object-level change detection methods to solve the problem of registration errors, to which pixel-level change detection is sensitive; the phenomenon of noise is serious, object-level change detection is greatly affected by segmentation parameters, and the process is cumbersome but still affected by the selection of scale parameters [10]. At the same time, although high-resolution remote sensing images can present more detailed information, the separability between changed and unchanged areas is reduced, and the loose spatial dependency between ground objects and highly cluttered spatial structures increases the difficulty of extraction [11]. In addition, the geometric position difference caused by different shooting angles of the sensor is a non-negligible interference factor. These factors weaken the separability of spectral information and make it more difficult to detect changes in high-resolution images. Therefore, it is of great scientific significance and application value to carry out accurate change detection algorithm research for high-resolution images.

Given the rise of artificial intelligence, deep learning algorithms with stronger image semantic feature extraction abilities have been gradually introduced into the research of high-resolution remote sensing image interpretation. The remote sensing image CD method based on deep learning can directly learn the change features from bi-temporal, multi-temporal, or even time-series remote sensing images, segment the image through the change features to obtain the change map, and the learned features have strong robustness. A Siamese neural network extracts features in the same way through multiple inputs for comparison between images and is often used in change detection tasks. For example, Zhan and Zhang proposed a method based on a Siamese convolutional neural network and applied it to change detection in optical aerial images [12,13]. Hughes et al. proposed a method based on a pseudo-Siamese convolutional neural network and applied it to change detection of SAR and optical images [14]. Daudt et al. took the U-Net structure as the backbone extraction network and proposed a Siamese fully convolutional network for change detection [15]. In addition, Zhang et al. extracted the highly representative features of bi-temporal images by using a dual-stream structure and then input the extracted features into the deep supervision difference recognition network for remote sensing image change detection [16]. Chen et al. proposed a dual-attention Siamese network, captured the long-term dependency through the dual-attention mechanism to obtain a more distinctive feature representation, then adjusted the weight of the unchanged feature in the training process by using the weighted double-margin contrastive loss, which achieved reasonable results on a public dataset [17]. Dong et al. proposed a multi-scale context aggregation network (MSCANet) to aggregate multi-scale context information using a scale-aware feature pyramid module (FPM) and discriminative feature representation learning with channel-spatial attention module to improve recognition performance [18]. Fang et al. proposed a dense connection Siamese network to reduce the loss of deep localization information of neural network through compact information transmission and used an ensemble attention module to extract representative features of different semantic levels for remote sensing image change detection [19]. Because existing methods fail to predict the edges and preserve the shape of the changed area from bi-temporal images, Basavaraju et al. introduced a network based on an encoder-decoder architecture (UCDNet) that uses improved residual connections and a new spatial pyramid pooling (NSPP) block to obtain better prediction results while preserving the shape of changing regions [20]. Chen and Lu et al. proposed Siamese-AUNet by combining a Siamese network, attention mechanism and U-Net for the detection of weakly changing objects, and the representation ability of weak features is improved by combining a non-local attention module and convolutional block attention module (CBAM), then the ASPP module is used to improve the detection effect of multi-scale change features [21].

At present, deep learning change detection methods still have some problems, such as difficulty detecting small-scale changes, loss of information in the process of feature

encoding and decoding, integrity of detection results, unbalanced sample ratio, and poor detection effects on non-public datasets. To solve these problems, we proposed a multi-scale attention Siamese network (MASNet) and tested its performance in change detection of high-resolution dataset in Yunnan Province, China (YNCD) in this paper. MASNet firstly employed a ResNet-50 backbone to capture bi-temporal features. Then, a multi-scale attention (MSA) module was utilized to model and generate multi-scale features. Finally, a prediction module was applied to obtain change maps by threshold and bilinear interpolation. In addition, by improving the contrast loss function, we could improve the imbalance of the samples and highlight the changed information. Considering that the result accuracy and model robustness of the deep learning change detection algorithm are greatly affected by the data samples, in order to carry out research on the change detection of remote sensing images in the experimental area, this research obtains change detection samples of the Yunnan experimental area, and adds a small amount of this data to the public dataset, then discusses the feasibility of the dataset expansion method by comparing the change detection result accuracy between the experimental area samples with and without added data. On this basis, the effectiveness of the MASNet model was verified in comparative experiments and ablation studies.

2. Materials and Methods

2.1. Change Detection Model

2.1.1. Siamese Network

As shown in Figure 1, a Siamese network is different from a general convolutional network. The Siamese network maps two different inputs into vectors, calculates the distance between different vectors through two stream networks with the same weight. It then compares the similarity between the bi-temporal images. These two types of networks have the same weights, update parameters, and share weights at the same time. After several convolutions, the feature maps carrying semantic information in each branch are converted into feature vectors. The feature vectors are filtered and fused to output decision information.

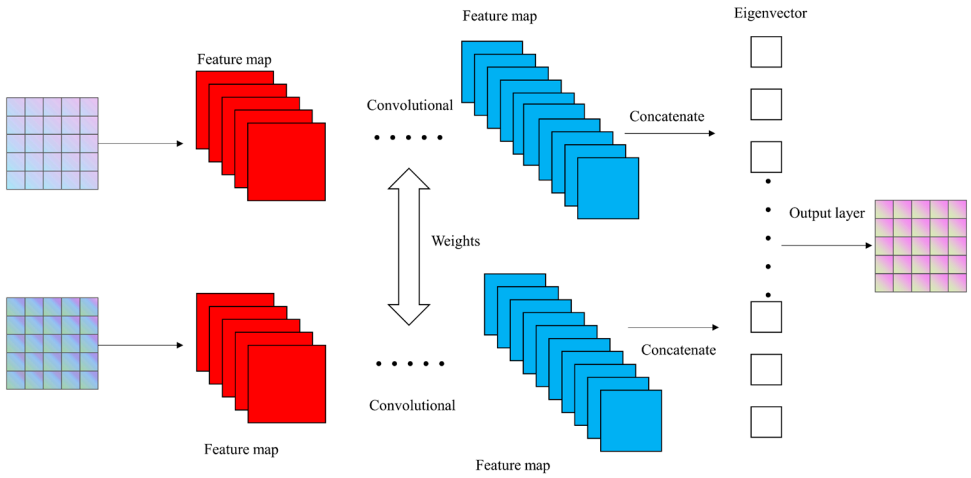

Figure 1. Siamese network architecture.

2.1.2. Attention Mechanism

The attention mechanism originated from the study of human vision. When people observe an object, they selectively pay attention to part of all the available information, ignoring other visible information. This mechanism is usually called the attention mechanism. The attention mechanism in deep learning was first used in natural language processing (NLP). In 2014, Mnih et al. used the attention mechanism for the first time in the field of image processing [22]. Its essence was to learn the weight distribution in the image, assign different weights to different important areas. The attention mechanism suppresses the learning of such features by reducing the weight of features unrelated to the target, and

at the same time increases the weight of relevant features to strengthen the learning of these features.

The position attention module (PAM) was proposed by Fu et al. [23], to capture rich global relationships between pixels in spatial locations in scene segmentation tasks to capture more discriminative features. Based on this understanding, we introduced PAM into MASNet. Its structure is shown in Figure 2:

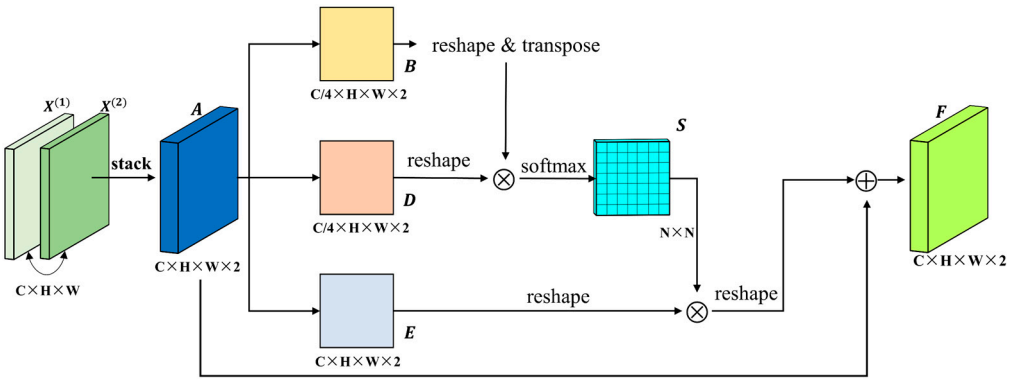

Figure 2. Structure of PAM model.

First, we used the stack of feature maps $X^{(1)}, X^{(2)}$ extracted from two different phase images to generate the feature tensor A and input it into the PAM model. The PAM model obtained two new feature layers through a convolution layer with batch normalization and a ReLU layer. To reduce the number of feature dimensions, the number of channels was changed to a quarter of the original, and then the two new feature layer dimensions are converted to $R \in (C \times N)$, where C refers the number of channels of the feature map, $N \in H \times W \times 2$, and H, W are the width and height of the feature map, respectively. Then, the softmax function was used to calculate matrix multiplication between B, D to obtain the similarity matrix S, which can be denoted as

$$S = softmax\left(\frac{B^T \cdot D}{\frac{1}{2}\sqrt{C}}\right) \tag{1}$$

The same convolution layer was used to generate the feature map E. We multiplied feature map E by similarity matrix S, and added the result to the original feature map A to obtain the result matrix F containing spatial position features. This can be denoted as

$$F = (E \cdot S) + A \tag{2}$$

2.1.3. Multi-Scale Attention Siamese Network

According to the principle of the neural network and attention mechanism, we introduced a multi-scale attention mechanism into the Siamese network to detect changes between image in the experimental area. As shown in Figure 3, MASNet contains three parts: Siamese CNN feature extractor, multi-scale attention module, and prediction module. Detailed information about each part are as follows.

The MASNet feature extraction module uses Siamese ResNet-50 without the global pooling layer and the fully connected layer as the backbone network to extract bi-temporal features. This includes a convolutional layer, maximum pooling layer, and 4 residual blocks (ResBlock). Each residual block contains {3,4,6,3} 3×3 convolutional layers, and the features in each ResBlock are fused by adding the upper- and lower-level feature information before being input to the ReLU layer. The number of output channels of each ResBlock is 256, 512, 1024, and 2048.

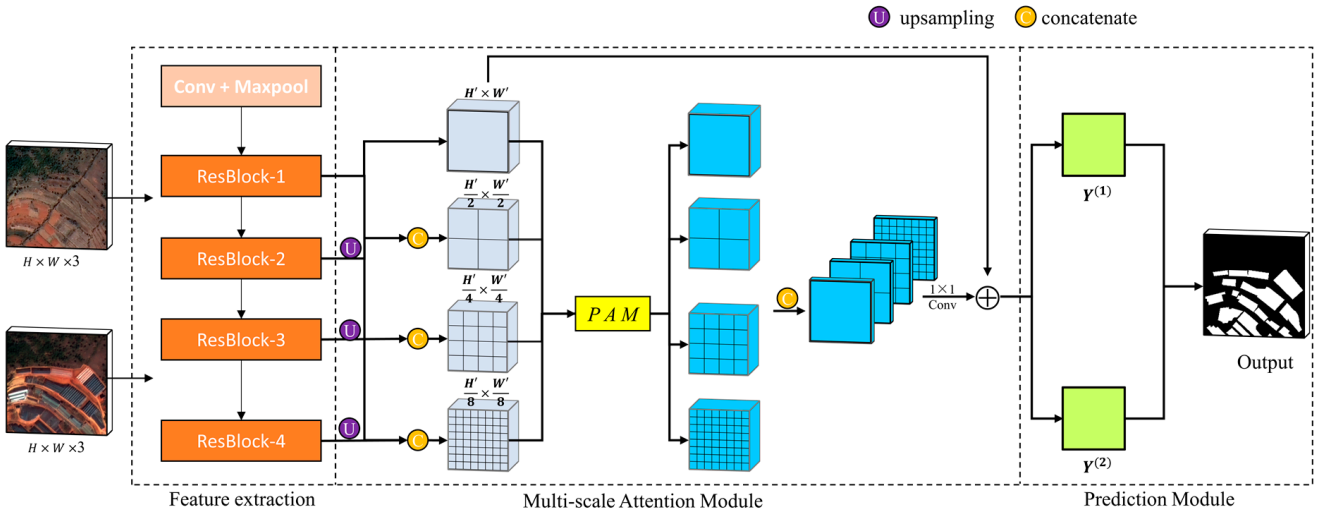

Figure 3. Illustration of multi-scale attention Siamese network architecture.

Feature maps of different scales contain semantic information of different scales. Large-scale feature maps have more global semantic information, and small-scale feature maps tend to highlight local semantic information. The semantic information in feature maps of different scales is fused to enhance the feature information extraction capability. According to Figure 3, the multi-scale attention module in MASNet applies PAM to the features of four different scales extracted by the feature extraction module. The attention mechanism is often used in the final feature map extracted by the feature extraction module. Since the continuous convolution process causes a loss of information in the image, we changed it to act on the feature map generated by the feature extraction module at different stages. PAM was directly applied to the feature map with the size of $H' \times W'$ extracted by ResBlock-1. In this paper, the size of the original image after one convolution and maximum pooling is $H' = W' = 128$. Other ResBlock feature extraction results are up-sampled and concatenated with low-level features to generate multi-scale attention features through PAM to obtain local feature representation at different scales. The process in MSA can be expressed as

$$Fea_i = Concat\left(\sum_{n=1}^{i} ResBlock - n\right), i = 1, 2, 3, 4 \tag{3}$$

The attention maps of different scales generated by PAM are concatenated with the original feature maps after passing through 1×1 Conv, where the sizes of the feature maps are $\left\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\right\} H' \times W'$ to generate the to-be-predicted results containing feature information of different scales.

The prediction module interpolates the multi-scale features extracted by the feature extraction module and the multi-scale attention module to the original image size using bilinear interpolation. It then calculates the Euclidean distance between feature pixels and obtains the final change detection result through threshold segmentation.

2.2. Loss Function

The traditional contrastive loss can be formulated as

$$CL = \sum_{i,j} \frac{1}{2}\left[(1 - y_{i,j}) \cdot D_{i,j}^2 + y_{i,j} \cdot max\left(m - D_{i,j}, 0\right)^2\right] \tag{4}$$

where $y_{i,j}$ represents the pixel value at the position of (i, j), 0 indicates that the pixel is unchanged, 1 indicates that the pixel has changed, $D_{i,j}$ indicates the distance between the pixels corresponding to position (i, j) on the bi-temporal image, and m is the threshold

value of the change feature. When the distance between the change features distributes in 0 to m, the loss value gradually decreases as $D_{i,j}$ increasing the distances among the changed pixels increased by optimizing the loss function to minimize the loss value. From Equation (4), it was found that when the distance value between unchanged image pairs is 0, the loss function is not affected. However, in practice, the image is affected by factors such as sensors and illumination. The distance value between unchanged image pairs is not 0, which leads to the unchanged image pairs also affecting the loss function. In addition, in the change detection task, the proportion of unchanged pixels is usually larger than that of changed pixels, and there is an imbalance between changed and unchanged samples. Referring to the CL designed by Chen [17], and Chen [24], we improve the contrastive loss function (ICL) as

$$ICL = \sum_{b,i,j} \frac{1}{2}\left[\frac{1}{N_u}\cdot\left(1 - M_{b,i,j}\right)\cdot max\left(D_{b,i,j} - m_u, 0\right)^2 + \frac{1}{N_c}\cdot M_{b,i,j}\cdot max\left(m_c - D_{b,i,j}, 0\right)^2\right] \quad (5)$$

Similar to Equation (4), in Equation (5), we added a threshold for unchanged pixels. When the distance between unchanged pixels is less than m_u, the loss function is not affected. The thresholds m_u, m_c representing unchanged and changed features are set to 0.2 and 2.0 in this paper, respectively. At the same time, considering the imbalance between the unchanged and the changed samples, a self-adjusting weight coefficient $\frac{1}{N_u}, \frac{1}{N_c}$ is introduced according to the ratio of the number of unchanged and changed pixels in different images of the current training batch to reduce the impact of the unchanged region on the changed region. This can be defined as follows:

$$N_u = \frac{\sum_{i=1}^{B} n_{u_i}}{B} \quad N_c = \frac{\sum_{i=1}^{B} n_{c_i}}{B} \quad (6)$$

B represents the batch size, and n_u, n_c represents the number of unchanged and changed pixels, respectively. At the same time, the training idea of [24], was introduced into Formula (5), where M represents a batch of change labels, and b represents a pair of images in a batch. The previous training from a single image was changed to batch image training, which made the training process smoother. The loss curve did not shock greatly and was consistent with the self-adjusting weight coefficients in terms of batch size in this paper.

3. Dataset

3.1. CDD

The Change Detection Dataset (CDD) is an open-source multi-type change detection dataset [25], that covers remote sensing images of the same area with seasonal changes. CDD contains 11 pairs of 0.3–1 m high-resolution images, of which there are 7 pairs of images with a size of 4725 × 2200 that change with seasons and 4 pairs of images with a size of 1900 × 1000 that change with seasons without labels. In the process of making labels, only the appearance or disappearance of objects is considered to have changed, while ignoring the changes caused by seasonal differences, brightness, and other factors. In [25], the author processed the original data to generate the dataset consists of 16,000 pairs of samples of images with 10,000 pairs for training, 3000 pairs for validation, and 3000 pairs for testing. Examples of CDD are displayed in Figure 4.

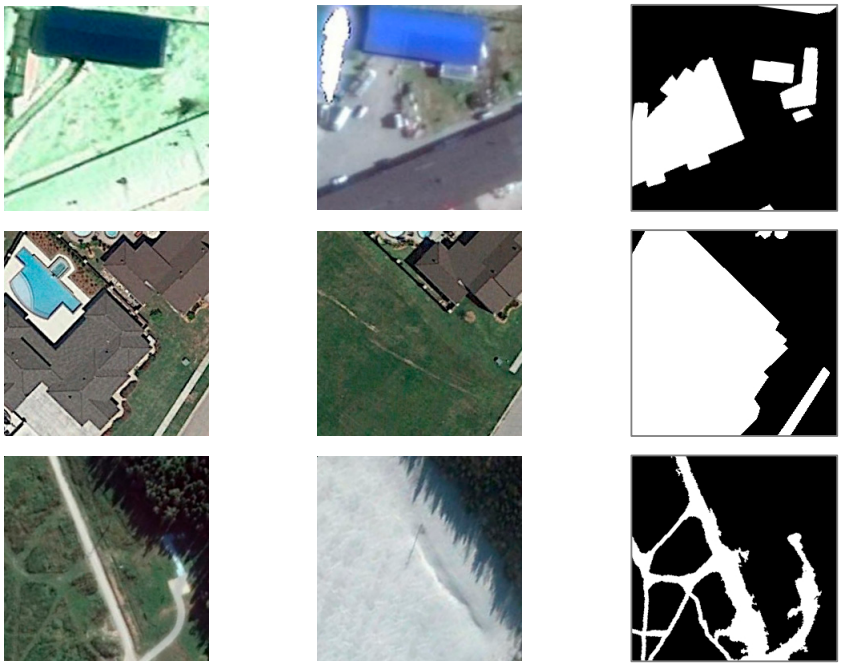

Figure 4. Part samples in CDD.

3.2. YNCD

In this paper, high-resolution satellite images obtained by the Gaofen-2 PMS sensor on 11 March 2015 and 25 February 2021, are used to generate image pairs with a resolution of 1 m after preprocessing, registration, and fusion. The data cover part areas of Kunming and Qujing city in Yunnan Province, China. According to a size of 256 × 256 pixels, the original image was cropped without overlapping to obtain 1540 pairs of change samples. The changed information was marked based on the visual interpretation of artificial targets such as protected agriculture, buildings, roads, mining areas, and small targets in the test area. We used the LabelMe tool for labeling, which is an image labeling tool developed by the Massachusetts Institute of Technology (MIT) Computer Science and Artificial Intelligence Laboratory (CSAIL). In order to make the outline and range of the change area more accurate, we marked the changes on the bi-temporal images respectively, and then combined them to obtain the bi-temporal change labels. We used 255 and 0 for changed and unchanged pixels, respectively. Table 1 gives a brief introduction of YNCD. Among them, the changes caused by seasonal differences, brightness, and other factors were ignored in the process of change information labeling. Examples of YNCD are displayed in Figure 5. The change detection data in the Yunnan experimental area were divided to three parts: 1000 pairs of images used as training samples; 300 pairs of images used as validation samples; and 240 pairs of images used as testing samples. The data from the Yunnan experimental area (YNCD) and CDD data were combined to generate the dataset used for change detection analysis, in which the proportions of the data from the Yunnan experimental area in the training, validation, and test sets were 10%, 10%, and 8%, respectively.

Table 1. A brief introduction to YNCD dataset.

Attribute	Value
Total Image Pairs	1540
Total Changed Pixels	8,000,699
Total Unchanged Pixels	92,924,741
Image Size	256 × 256
Image Resolution	1 m
Time Interval	6 years
Type	RGB image

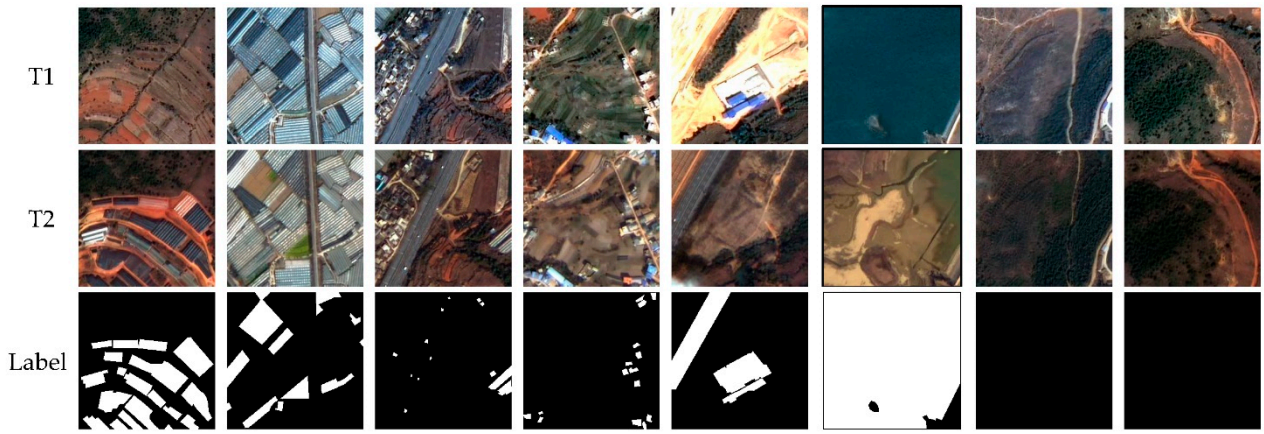

Figure 5. Change detection samples in the experimental area of Yunnan Province. Each column represents one sample. Columns 1 and 2 show the protected agriculture change; 3–6 show the building, small targets, roads, and water change; and 7, 8 display samples of no change.

4. Experiments and Results

4.1. Parameters Setting

The code is implemented using the PyTorch framework. During the training process, a batch size of four and a learning rate of 0.001 were adopted for all model training using an Adam optimizer. The learning rate was adjusted through a dynamic learning rate decay strategy based on the deviation of the validation set. The training process lasted for 150 epochs, while data augmentation strategies including flipping, mirroring, and random rotation were randomly applied to the training set to avoid overfitting. We used both YNCD and CDD for training and validation. In the testing process, the test set of YNCD was used to evaluate the accuracy of the change detection results.

Furthermore, owing to the selection of the values of m_u and m_c having a great influence on the results of the network, lots of experiments were designed to find the values of m_u and m_c with the best performance. We set up m_u = {0.0,0.1,0.2,0.3,0.4}, m_c = {1.8,2.0,2.2}, and trained through the pairwise combinations between them to seek the optimal match on the test set. Figure 6 shows the performance of a MASNet network with different m_u and m_c values. It is not difficult to see that when m_u and m_c are 0.2 and 2.0 respectively, the performance of MASNet is the best.

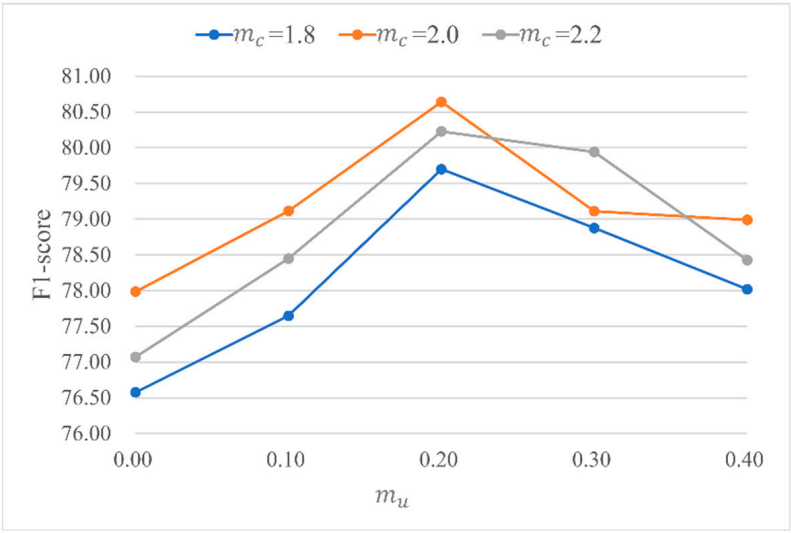

Figure 6. Influence of different values of m_u and m_c on the performance of MASNet.

4.2. Comparative Experiment

4.2.1. Evaluation Metrics

To verify the effectiveness of the proposed method, four common metrics were selected for accuracy assessment: precision (P), recall (R), F1-score ($F1$), and overall accuracy (OA). Their calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

$$OA = \frac{TP + TN}{TP + TN + FN + FP} \tag{10}$$

where TP indicates the correct number of changed pixels detected, TN indicates the correct number of unchanged pixels detected, FP indicates the error number of changed pixels detected, and FN indicates the error number of unchanged pixels detected. P indicates the ratio of the correct number of changed pixels detected to the total number of changed pixels detected. A larger P indicates less probability of false prediction. R indicates the ratio of the detected correct number of changed pixels to the total number of actual changed pixels. A larger R indicates less probability of missed detection. $F1$ considers P and R comprehensively and stands for the overall performance. OA indicates the ratio of the number of pixels detected correctly to the total number of pixels.

4.2.2. Comparative Analysis of Results of MASNet with and without YNCD

To evaluate the influence of the presence of YNCD samples in the training samples on the change detection results and to verify the feasibility of expanding the experimental area samples to the public dataset to improve the detection accuracy of change detection, the model detection results with and without experimental area samples are compared and analyzed in this paper. The MASNet model was trained by using CDD and part of YNCD samples, and change detection and analysis are carried out for the test samples of YNCD. The P, R, $F1$, and OA results are compared in Table 2, and prediction change maps of MASNet are shown in Figure 7.

Table 2. Comparison of change detection index with or without YNCD samples (%).

Datasets	*OA*	*P*	*R*	*F1*
CDD	90.32	34.61	19.49	24.94
CDD + YNCD	95.34	79.78	81.52	80.64

According to the evaluation indicators in Table 2, the results of model training using only the CDD and change detection in the experimental area were poor in accuracy, but the OA indicators were high. The reason for this result is that the number of unchanged pixels in YNCD accounted for a relatively high proportion, resulting in a large number of OA values. After adding the YNCD data, the evaluation indicators of P, R, $F1$ increased by 45.17%, 62.03%, and 55.7%. According to Figure 7, the model trained with CDD could detect the changes in a small part of the 3rd and 5th columns but hardly detected the changes in the other groups of images correctly. This phenomenon shows that the change detection model is greatly affected by the training samples. Compared with the label, the overall result of the change detection obtained by the training model after adding the YNCD was better, the detection area was relatively complete, and the phenomenon of missed detection was reduced. This indicates that even if a small amount of sample of the experimental area is added to the public dataset, the model can also obtain better results.

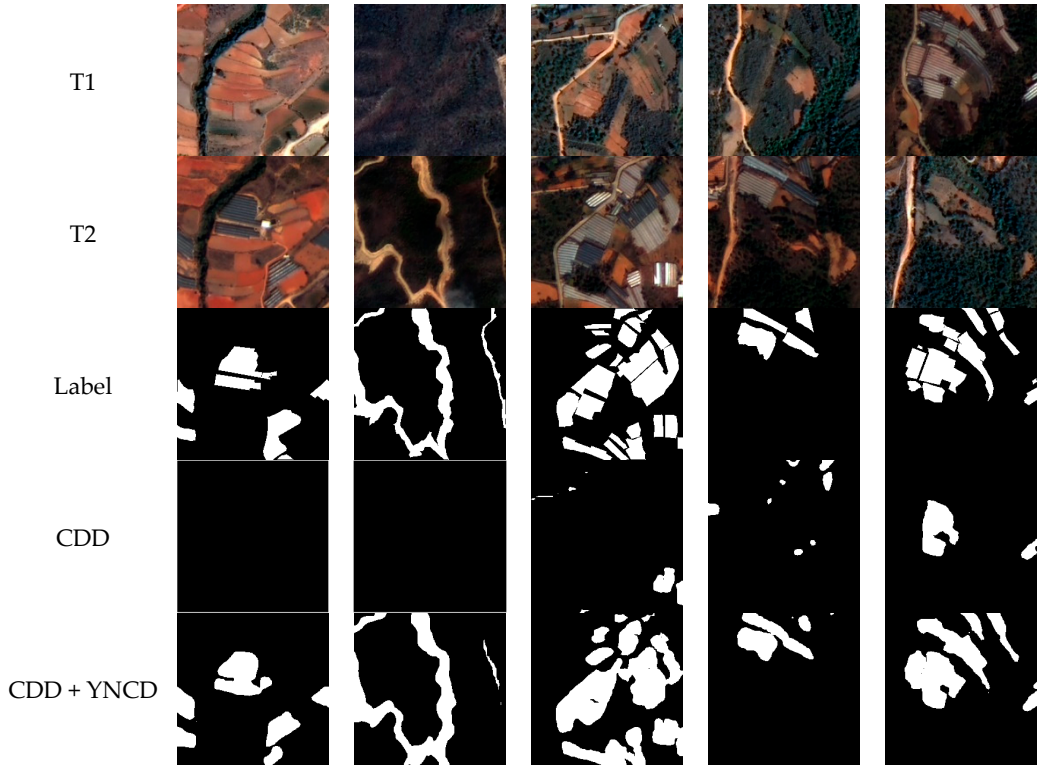

Figure 7. Comparison of change detection results with or without YNCD in model training.

4.2.3. Comparison Methods and Result Comparisons

Six methods for bi-temporal change detection were employed in our experiments for comparison.

(1) STANet (Spatial-Temporal Attention Network) [24]: STANet utilizes a weight-sharing Siamese deep semantic segmentation network to generate two independent features and designs a spatial-temporal attention mechanism that captures the rich global spatial-temporal relationships between pixels in the whole spatial-temporal space and generates more discriminative features by calculating the attention weights of any two pixels at different times and locations. Finally, the metric learning method is used to calculate the distance between the two features and generate prediction maps.

(2) MANet (Multi-scale Attention Net) [26]: The MANet network designs two modules: the Position-wise Attention Block (PAB) and the Multi-Scale Fusion Attention Block (MFAB). PAB analyzes the interdependencies of modeled features in the spatial dimension, thereby capturing the spatial dependencies between pixels in a global view. MFAB captures the channel dependencies between any feature maps through multi-scale semantic feature fusion.

(3) PAN (Pyramid Attention Networks) [27]: PAN method combines an attention mechanism with a spatial pyramid, performs a spatial pyramid attention structure on high-level feature output, and combines global pooling to learn better feature representation. A Global Attention Upsample module is introduced on each decoder layer to serve as the global contextual information for the localization details of the low-level feature selection category.

(4) FC-EF (Fully Convolutional Early Fusion) [15]: FC-EF is based on U-Net model and EF strategy. It concatenates the bi-temporal images before passing them through the network and uses the skip connection structure to fuse the low- and high-level features.

(5) FC-Siam-Conc (Fully Convolutional Siamese-Concatenation) [15]: With a similar structure to the feature extraction module in MASNet, a Siamese network model combines three features from the two encoder branches and corresponding layers of the decoder. The graph performs skip connections to supplement the deeper, more

abstract, and less localized information with low-level spatial detail information, resulting in more accurate boundary predictions in the output image.

(6) FC-Siam-Diff (Fully Convolutional Siamese-Difference) [15]: The difference between FC-Siam-Diff and FC-Siam-Conc is that different temporal features are not processed in channel stacking, but the absolute difference of bi-temporal image features in the encoder is combined with the features of the decoder through skip connections in FC-Siam-Diff method.

The CDD combined with part of YNCD samples was used for comparative experimental analysis, and the comparative experimental accuracy evaluation was conducted only for the YNCD data. The prediction change maps of several methods are illustrated in Figure 8, and the evaluation results are given in Table 3.

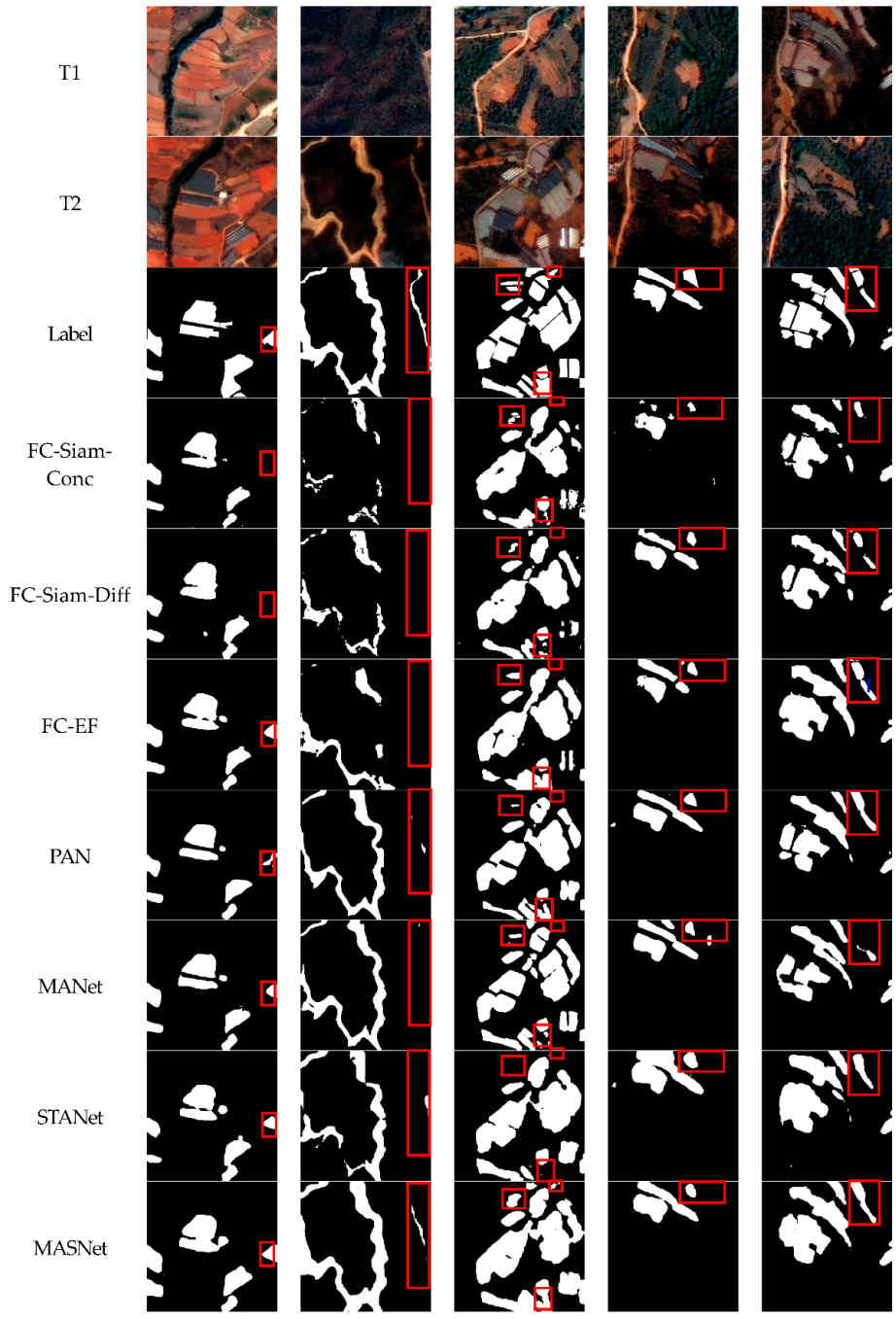

Figure 8. Comparison of change detection results of different models.

Table 3. Comparison of change detection indices for different models (%).

Method	*P*	*R*	*F*1
FC-Siam-Conc	64.94	53.78	58.84
FC-Siam-Diff	63.82	60.08	61.89
FC-EF	70.31	60.65	65.12
PAN	73.14	66.13	69.46
MANet	74.62	73.62	74.12
STANet	78.46	77.00	77.72
MASNet	79.78	81.52	80.64

From the indicators in Table 3, the *P*, *R* and *F*1 of MASNet are 79.78%, 81.52%, and 80.64%, respectively, which are significantly improved compared to other models for many indicators. Compared with the STANet model with the best detection effect in the comparison model, the indicators increased by 1.32%, 4.52%, and 2.92%, respectively, which proves that MASNet has the best performance in the experimental area. According to Figure 8, especially for the results in the red box area, the advantages and disadvantages of the detection results of different methods are shown. FC-Siam-Conc has the worst change detection results because its encoding and decoding process is relatively simple and no attention mechanism is introduced, so the phenomenon of false detection and missed detection is serious. FC-Siam-Conc has the same problem, but FC-Siam-Diff and FC-EF work better than FC-Siam-Conc, with *F*1 of 61.89% and 65.12%, respectively. Compared with FC-Siam-Conc, due to PAN and MANet introducing pyramid attention and multi-scale attention mechanisms, respectively, the detection results greatly improve, but the change region detection is still incomplete. STANet adopts the strategy of contrastive loss and spatial-temporal attention and obtains good change detection results, but there are still problems for small-scale change detection. In this paper, by improving the contrastive loss to enhance the characteristics of changing regions, reducing the impact of sample imbalance and adopting a multi-scale attention mechanism to improve the change detection ability of regions at different scales, the performance of the MASNet model was improved.

4.3. Ablation Study

In this part, we conducted an ablation study to verify the effectiveness of the ICL and MSA introduced in MASNet. "Test ①" used only contrastive loss and MSA for model training. "Test ②" used only the ICL to optimize the network. "Test ③" used both MSA and ICL to optimize the network. Figure 9 provides visualized comparisons of the ablation and evaluation results listed in Table 4.

Table 4. Ablation study evaluation results of loss function and attention module (%).

	ICL	MSA	*OA*	*P*	*R*	*F*1
①		✓	94.25	70.46	74.03	72.20
②	✓		94.61	71.78	78.26	74.88
③	✓	✓	95.34	79.78	81.52	80.64

From the comparison results of various indicators in Table 4, the improved contrastive loss and the multi-scale attention module comprehensively improved the performance of the model, and the indicators of *F*1 increased by 5.76% and 8.44%, respectively. In Figure 9, small-scale areas were missed in the change detection results of the first and third columns for "Tests ①" and "Test ②", and the detection areas were incomplete and false in the results of the other columns. The improvement in model performance can be attributed to the following aspects. Firstly, reducing the false change information extraction, the phenomenon of incomplete detection area and false detection is improved by improving the contrastive loss. Secondly, by setting self-adjusting weight coefficients to enhance the extraction of changing features, the problem of imbalance between the categories of invariant pixels and changing pixels is alleviated. Finally, the multi-scale attention module can obtain the feature information at different scales. The detection

effect of adding the multi-scale attention module alone in "Test ①" is not ideal. This phenomenon may occur because the pseudo-change information has a great negative impact on the model training during the training process. However, from the change detection results of "Test ③", on the basis of improving the contrastive loss, adding the multi-scale attention module enhances the detection effect of small-scale regions, which illustrates the effectiveness of the multi-scale attention module.

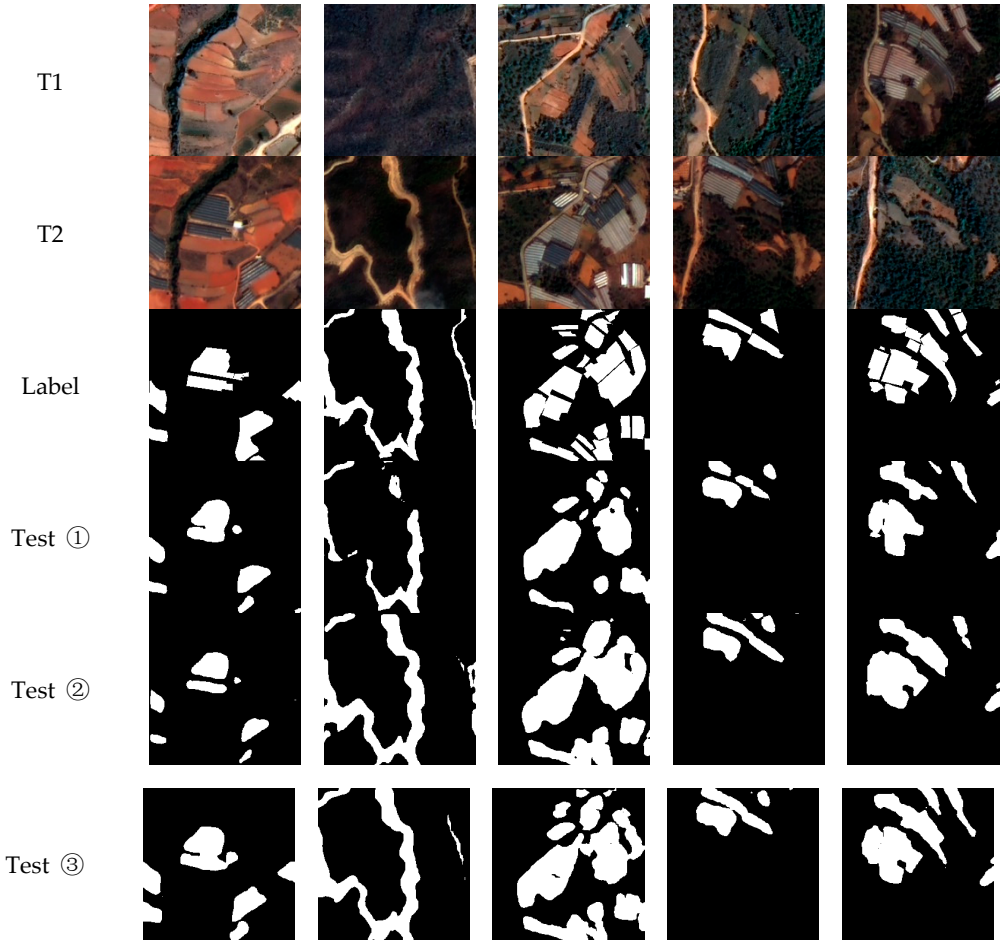

Figure 9. Comparison of partial change detection results of ablation study.

5. Conclusions

To solve the problems of incomplete detection of change regions and missed detection of small-scale areas in remote sensing image change detection, a multi-scale attention Siamese network (MASNet) was constructed based on a Siamese network and attention mechanism. Using the remote sensing image data of the Gaofen-2 PMS sensor, a high-resolution change detection dataset of some regions in Yunnan was created and combined with the CDD to carry out research on the change detection method of the experimental area.

We set up comparative experiments and an ablation study to verify the performance of MASNet. The following conclusions are drawn: (1) The detection results of the change detection model are greatly affected by the samples. Adding a small amount of experimental area data to the model training dataset can improve the experimental area change detection results and save on manual marking costs to a certain extent. This has important reference and application value for solving the change detection task when samples are lacking in the test area. (2) MASNet can suppress the information of the unchanged region by improving the loss function, highlighting the information of the changed region, strengthening the learning of the features of the changed region, and optimizing the change detection results. (3) The multi-scale attention mechanism can enhance the information extraction

of different scale regions and improve the model detection performance. (4) Compared with STANet, FC-Siam-Conc, FC-EF, FC-Siam-Diff based on the Siamese architecture, and MANet and PAN models based on multi-scale attention, the MASNet model has the best overall performance and change detection results. Compared with other change detection models, MASNet improved the F1-score and recall by at least 2.92% and 4.52%, respectively. However, it can also be seen from the change detection results that the MASNet model is ineffective in detecting the boundary and internal details of the changed area, and further research work will be carried out on this issue in the future.

Author Contributions: Conceptualization, S.Z. and J.L.; methodology, G.Z. and S.Z.; software, J.L.; validation, G.Z. and Y.X.; dataset production and curation, Y.G. and J.L.; writing—original draft preparation, J.L.; writing—review and editing, S.Z. and Y.X.; project administration, G.Z.; funding acquisition, S.Z. and Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (42171101, 41871028).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge the China Centre for Resources Satellite Data and Application for the provision of Gaofen-2 PMS data, and the Foundation for financial supports of this work.

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

References

1. Tong, G.F.; Li, Y.; Ding, W.L.; Yue, X.Y. Review of remote sensing image change detection. *J. Image Graph.* **2015**, *20*, 1561–1571.
2. 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]
3. Celik, T. Unsupervised change detection in satellite images using principal component analysis and k -means clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [CrossRef]
4. Nielsen, A.A.; Conradsen, K.; Simpson, J.J. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sens. Environ.* **1998**, *64*, 1–19. [CrossRef]
5. Wu, C.; Du, B.; Zhang, L.P. Slow feature analysis for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 2858–2874. [CrossRef]
6. Qin, Y.; Niu, Z.; Chen, F.; Li, B.; Ban, Y. Object-based land cover change detection for cross-sensor images. *Int. J. Remote Sens.* **2013**, *34*, 6723–6737. [CrossRef]
7. Ma, L.; Li, M.; Blaschke, T.; Ma, X.; Tiede, D.; Cheng, L.; Chen, Z.; Chen, D. Object-based change detection in urban areas: The effects of segmentation strategy, scale, and feature space on unsupervised methods. *Remote Sens.* **2016**, *8*, 761. [CrossRef]
8. Zhang, Y.J.; Peng, D.F.; Huang, X. Object-based change detection for VHR images based on multiscale uncertainty analysis. *IEEE Geosci. Remote Sens. Lett.* **2017**, *15*, 13–17. [CrossRef]
9. Zhang, C.S.; Li, G.J.; Cui, W.H. High-resolution remote sensing image change detection by statistical-object-based method. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2440–2447. [CrossRef]
10. Wu, R.J.; He, X.F.; Wang, J. Coastal wetlands change detection combining pixel-based and object-based methods. *J. Geo-Inf. Sci.* **2020**, *22*, 2078–2087.
11. Zhang, X.L.; Chen, X.W.; Li, F.; Yang, T. Change detection method for high resolution remote sensing images using deep learning. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 999–1008.
12. Zhan, Y.; Fu, K.; Yan, M.; Sun, X.; Wang, H.; Qiu, X. Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1845–1849. [CrossRef]
13. Zhang, Z.; Vosselman, G.; Gerke, M.; Tuia, D.; Yang, M.Y. Change detection between multimodal remote sensing data using siamese CNN. *arXiv* **2018**, arXiv:1807.09562.
14. Hughes, L.H.; Schmitt, M.; Mou, L.; Wang, Y.; Zhu, X.X. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 784–788. [CrossRef]
15. 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: New York, NY, USA, 2018; pp. 4063–4067.
16. 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]

17. 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]
18. Dong, J.; Zhao, W.F.; Wang, S. Multiscale context aggregation network for building change detection using high resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
19. 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]
20. Basavaraju, K.S.; Sravya, N.; Lal, S.; Nalini, J.; Reddy, C.S.; Dell'Acqua, F. UCDnet: A deep learning model for urban change detection from bi-temporal multispectral sentinel-2 satellite images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–10. [CrossRef]
21. Chen, T.; Lu, Z.; Yang, Y.; Zhang, Y.; Du, B.; Plaza, A. A Siamese Network Based U-Net for Change Detection in High Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2357–2369. [CrossRef]
22. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. *arXiv* **2014**, arXiv:1406.6247.
23. Fu, J.; Liu, J.; Tian, H.J.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: New York, NY, USA, 2019; pp. 3146–3154.
24. Chen, H.; Shi, Z.W. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
25. Lebedev, M.A.; Vizilter, Y.V.; Vygolov, O.V.; 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]
26. Fan, T.L.; Wang, G.L.; Li, Y.; Wang, H. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access* **2020**, *8*, 179656–179665. [CrossRef]
27. Li, H.C.; Xiong, P.F.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.