remote sensing

MDPI

Article

A Deeply Supervised Attentive High-Resolution Network for Change Detection in Remote Sensing Images

Jinming Wu [1,2], Chunhui Xie [1,3], Zuxi Zhang [1,2] and Yongxin Zhu [1,*]

[1] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
* Correspondence: zhuyonxin@sari.ac.cn

Abstract: Change detection (CD) is a crucial task in remote sensing (RS) to distinguish surface changes from bitemporal images. Recently, deep learning (DL) based methods have achieved remarkable success for CD. However, the existing methods lack robustness to various kinds of changes in RS images, which suffered from problems of feature misalignment and inefficient supervision. In this paper, a deeply supervised attentive high-resolution network (DSAHRNet) is proposed for remote sensing image change detection. First, we design a spatial-channel attention module to decode change information from bitemporal features. The attention module is able to model spatial-wise and channel-wise contexts. Second, to reduce feature misalignment, the extracted features are refined by stacked convolutional blocks in parallel. Finally, a novel deeply supervised module is introduced to generate more discriminative features. Extensive experimental results on three challenging benchmark datasets demonstrate that the proposed DSAHRNet outperforms other state-of-the-art methods, and achieves a great trade-off between performance and complexity.

Keywords: change detection; convolutional neural network; feature fusion; metric learning; attention mechanism

Citation: Wu, J.; Xie, C.; Zhang, Z.; Zhu, Y. A Deeply Supervised Attentive High-Resolution Network for Change Detection in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 45. https://doi.org/10.3390/rs15010045

Academic Editors: Lizhe Wang, Xiaodong Zhang, Jining Yan and Guanzhou Chen

Received: 27 October 2022
Revised: 8 December 2022
Accepted: 20 December 2022
Published: 22 December 2022

1. Introduction

Change detection (CD) is the process of identifying differences in RS images at different times in the same geographical location [1]. It has been widely applied in such diverse disciplines as urbanization monitoring [2,3], environmental monitoring [4–7], and disaster assessment [8,9]. With the rapid development of Earth observation techniques including WorldView, GF, and Sentinel, huge amounts of RS datasets [10–15] from various sensors are available, which has created new demands on remote sensing change detection. Therefore, to exploit multitemporal information from RS data, a lot of efforts on developing new effective CD methods have been made in the field of remote sensing.

Over the past few decades, many remarkable change detection methods have been proposed. These CD methods can broadly be divided into two types: traditional and deep learning (DL) based. Most traditional CD methods detect changes by image mathematical and statistical techniques, including change vector analysis [16], principal component analysis [17], slow feature analysis [18], and multivariate alteration detection [19,20]. Handcrafted features or learnable rules are carefully designed to compare representations of data. Moreover, other machine learning techniques have been applied in CD as classifiers, including support vector machine [21], spatial domain analysis [22], decision tree [23], and random forest [24].

Recently, deep learning techniques, especially deep convolutional neural networks (CNN), have achieved breakthroughs in many fields. Researchers have introduced DL (VGG [25], Unet [26], ResNet [27], et al.) and semi-supervised learning [28–32] techniques to improve performance for CD algorithms. These powerful CNN models can learn hierarchical features. The shallow convolutional layer has a smaller receptive field and

generates high-resolution low-level features, which are rich in spatial information such as contour and shape. The deep convolutional layer has a larger receptive field and generates low-resolution high-level semantic features. For change detection, whether a change has occurred can be determined by comparing high-level semantic features, and where a change has occurred can be located by combining low-level spatial features. Most available DL-based methods have a deep encoder–decoder architecture for CD. The encoder that is composed of stacked convolution layers learns multi-level deep features with rich spectral and spatial features from remote sensing images. The decoder fuses multi-level features from the encoder to obtain change features; then, CD is converted into a task of pixel-wise classification to identify the change of interest.

In terms of the CD encoder, the latest efforts have been focused on the refinement of spatial-temporal feature discrimination, through stacking more convolution layers [11,33–35], and applying attention mechanisms [36–38]. However, these methods bring the high complexity and cost of computation to CD algorithms. The trade-off between detection precision and computational complexity must be well considered. In terms of the CD decoder, most efforts have been made on the separability of interclass. Depending on the type of classifiers, there are mainly two types of DL-based CD decoders: metric-based [10,12,35,39–41] and classification-based [11,33,34,36,37,42,43]. Previous metric-based CD decoders detect changes by binarizing the distance map between bitemporal features. These metric-based methods introduce metric learning to learn a parameterized embedding space for bitemporal images, where changed pairs should be as far away as possible, while unchanged pairs should be as close as possible. Different from metric-based methods, classification-based decoders convert CD into a classification task by assigning two scores to each point of the image, and two scores indicate the possibility that belongs to the change/unchange class, respectively.

Although many remarkable CD methods have been proposed, there are still some shortcomings that can be summed up as follows: (1) Most existing methods generate high-resolution change features by fusing and upsampling low-scale features, and thus noise or misalignment may be injected. (2) Most existing loss functions designed for CD networks are not effective enough to supervise the training process. Semantic information of bitemporal images is not fully exploited and some implicit relationships between bitemporal features may be ignored. Thus, it is still challenging to extract bitemporal features efficiently and change information correctly for DL-based CD methods.

In view of the above-mentioned problems, we propose a deeply supervised attentive high-resolution network to improve performance for CD. We construct the network based on two demands. The first is that the network must fully exploit semantic relationships between bitemporal features, which makes the subsequent decoding process of change features easier. The second is that the network must enhance features relevant to change of interest, which allows the changed areas to be identified more precisely. To satisfy the first demand, we augment a Siamese semantic segmentation branch for bitemporal images, which is different from existing methods. In addition, we utilize the change labels to provide supervision signals for the semantic branch, and propose a deep supervision strategy to assist the optimization process of the network. The deep supervision strategy can effectively improve the separability of semantic features in the embedding space. To satisfy the second demand, we refine high-resolution semantic features hierarchically that are extracted by the encoder, and enhance change features by a specially designed spatial-channel attention module in the decoder. Varying from existing methods, large kernel convolution in the attention module is decomposed to capture long-range dependence. Then, attention-refined features are fed into stacked convolution blocks in parallel to reduce feature misalignment. In view of the above efforts, our network is capable of capturing high-resolution discriminative features to improve performance for various kinds of changes in RS images.

The main contributions made by this work can be summed up as follows:

(1) We propose a novel classification-based change detection network based on encoding-decoding structure, which introduces an attention mechanism into the change decoder for the purpose of extracting more discriminative features. Thereby, a robust change detection algorithm is implemented.
(2) We introduce a novel deeply supervised strategy for intermediate layers in the change decoder to minimize intermediate feature distances. A new change loss is defined for the training process of DSAHRNet to improve the network's performance.
(3) We evaluate our network by comparing it with eight state-of-the-art (SOTA) change detection networks on three benchmark datasets, and extensive comparative experiments reveal that the proposed network outperforms these recent SOTA change detection networks, yielding the highest F1 score of 92.11%, 96.94%, and 83.57%, respectively.

The rest of this paper is organized as follows: The related works of DL-based CD methods are reviewed in Section 2. Section 3 introduces the proposed method in detail. The settings and results of experiments are reported in Section 4. Section 5 discusses our proposed method. Finally, Section 6 draws the conclusion of our work.

2. Related Work

This section provides a brief review of recent change detection methods with deep learning techniques in remote sensing. Change detection is a critical task in the field of remote sensing, and many DL-based change detection techniques have been developed in recent years. According to the feature learning process for input bitemporal images, the DL-based change detection networks can be roughly classified into two categories: single-stream and double-stream.

Single-stream networks usually use a semantic segmentation network to distinguish changes from the concatenation or difference of two bitemporal images. Daudt et al. [43] proposed a fully convolutional early fusion network based on UNet [26], named FC-EF. FC-EF concatenates the bitemporal images along the channel before feeding them into the network. Liu et al. [44] integrated depthwise separable convolution [45] into UNet to achieve better performance for change detection. Peng et al. [46] used a modified UNet++ [47], which is a powerful segmentation network built with dense skip connections, to detect changed regions.

Double-stream networks are commonly constructed on the Siamese structure. The Siamese network with sharing weights extracts multi-level features from bitemporal images, then the output features are concatenated along the channel to classify or calculate distance maps by a certain metric. Therefore, the discrimination of features and the feature fusion strategy are critical to identifying changes. To improve the discriminative power of networks, most recent works adopt multi-level fusion strategies [33,34,42,48], deep supervision [10,11] and attention mechanisms [12,36–38,49,50]. Shi et al. [10] proposed a dual attentive convolutional Siamese network in which the convolutional block attention module (CBAM) [51] was integrated to make features more discriminative. Zhang et al. [11] proposed a difference discrimination network and multiple deeply supervised modules to enhance intermediate layers' learning ability. Chen et al. [34] proposed a feature constraint change detection network. A novel dual encoder–decoder backbone was designed to suppress background noise, and a self-supervised learning strategy was introduced to constrain feature learning. Xu et al. [42] proposed a multidirectional fusion pathway strategy to boost information paths in the network. Chen et al. [37] proposed a transformer-based Siamese network to model contexts within the spatial-temporal domain. Nevertheless, these networks are inefficient to reduce feature misalignment and may suffer from the problem of high computational complexity brought from stacked convolution layers or multi-level feature fusions.

Although the above-mentioned methods have achieved good performance, most of them are in a supervised learning manner with annotated data. For practical purposes, it is more attractive to design semi-supervised CD methods that require partially labeled data sets. Peng et al. [28] proposed a semi-supervised CD network based on a generative adversarial network. In the network, the discriminator was employed to distinguish whether a change map is from the unlabeled samples or the ground truth. The generator learned the information from both the labeled and unlabeled samples through entropy adversarial training. Bandara et al. [29] proposed a semi-supervised CD framework to enhance the performance of CD approaches, where consistency regulation was applied to leverage the information from unlabeled bitemporal images. Apart from these methods, Zheng et al. [31] presented an unsupervised building CD network with single-temporal supervised learning to bypass the problem of collecting pairwise labeled images. In this method, first, pseudo bitemporal image pairs were constructed by random permutation in mini-batch to provide change supervisory signals and change labels were generated by logical exclusive OR (xor) operation on bitemporal building labels. Next, the network was trained with an effective multi-task architecture for joint semantic segmentation and change detection on the constructed pseudo bitemporal images.

Inspired by previous works, we incorporate an attention mechanism and deep supervision into the CD network in our method. We propose a novel spatial-channel attention module to learn change features in a spatial-wise and channel-wise manner. We apply multiple deeply supervised modules to assist network training by maximizing the similarity between the changed (unchanged) pairs. As a result, a high-performance double-stream CD network is constructed in this paper.

3. Methodology

In this section, we provide details of the proposed network. First, the overall procedure of the proposed DSAHRNet is illustrated (see Figure 1). Then, we present details of our designed SCAM and change decoder. Finally, the loss function of the network is defined.

3.1. Overview

The overall procedure of the proposed network is illustrated in Figure 1. It consists of three parts: a feature encoder, a deep attentive high-resolution change decoder, and a classifier. We adopt a modified HRNet18 [52] as our encoder to extract multi-level features. Then, the encoded features are fed to the proposed change decoder with the aim of learning long-range differences between bitemporal features. The change decoder is comprised of hierarchical feature fusion modules and parallel convolution blocks (PCBs). In the process of decoding, extracted bitemporal features are fused in a bottom-up manner by SCAM to generate change features. Next, the bitemporal features and change features are refined by stacked convolutional blocks in parallel. In this way, the bitemporal images are mapped into features in an embedding space, which is represented semantic information of bitemporal images. Change features are refined to reduce misalignment. Finally, the classifier predicts a pixel-level change map according to the concatenation of the change features. In the training process, the change map is exploited to calculate the classification loss (Equation (8)) with labels, and the bitemporal features are deeply supervised by the contrastive loss (Equation (7)). The hybrid loss (Equation (12)) is employed to improve the convergence stability of the proposed network and bring better detection results.

The inference detail of our DSAHRNet is presented in Algorithm 1. Let $\{(I^1, I^2)\}$ represent a pair of bitemporal images, and y represents the ground truth. The inference and training process of DSAHRNet can be summarized as follows:

(1) First, the bitemporal images $\{(I^1, I^2)\}$ are input into the weight-sharing Siamese encoder, with each obtaining a group of multi-level features $F^i = \{f_1^i, f_2^i, f_3^i, f_4^i\}$, $i = 1, 2$.
(2) Next, the same level features are merged at the channel dimension to obtain change features, $c_i, i = 1, 2, 3, 4$, where SCAMs are applied hierarchically.

(3) Then, both bitemporal features F^1, F^2 and change features $\{c_i, i = 1, 2, 3, 4\}$ are refined by PCBs and SCAMs.
(4) After that, the classifier generates a change mask M from multi-level change features $\{c_i, i = 1, 2, 3, 4\}$. L_{BCE} and L_{Dice} are calculated between the predicted change mask M and the ground truth y by Equation (8) and Equation (9), while L_{BCL} is calculated for intermediate bitemporal features by Equation (7).
(5) Finally, the sum of L_{BCE}, L_{Dice} and L_{BCL} are back-propagated to optimize model weights.

Algorithm 1 Inference of DSAHRNet for change detection

Input : $\mathbf{I} = \{(\mathbf{I^1}, \mathbf{I^2})\}$(a pair of bitemporal images)
Output : **M** (a prediction change mask)
// step1: feature extraction by HRNet18
for i in {1,2} **do**
    $\mathbf{F}^i = [f_1^i, f_2^i, f_3^i, f_4^i]$ = HRNet18($\mathbf{I^i}$)
end
// step2: use SCAM to generate change features hierarchically
c_0=None
for i in {1,2,3,4} **do**
    c_i = SCAM(Concat(f_i^1, f_i^2, c_{i-1}))
end
// step3: refine features
for i in {1,2,3,4} **do**
    f_i^1, f_i^2, c_i = PCB(f_i^1, f_i^2, c_i)
end
for i in {1,2,3,4} **do**
    c_i = SCAM(Concat(f_i^1, f_i^2, c_{i-1}))
end
// step4: obtain change mask by the classifier
M = Classifier(Concat(c_1, c_2, c_3, c_4))

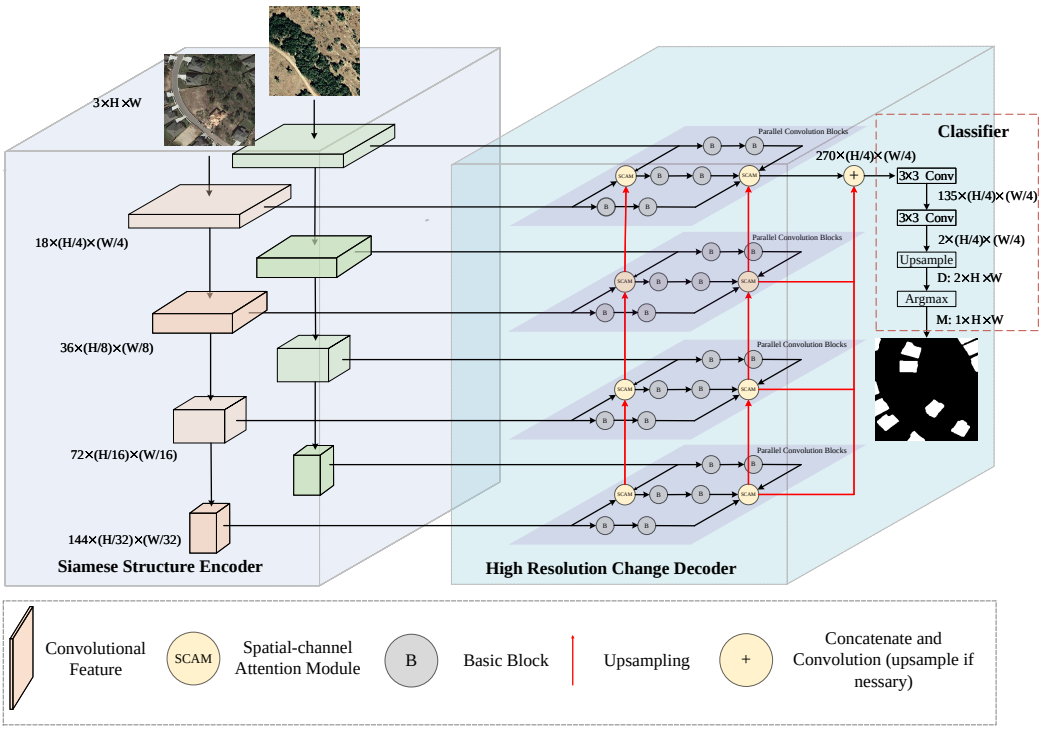

Figure 1. Overview of our DSAHRNet model.

3.2. Deep Attentive High-Resolution Change Decoder

The change decoder is designed to generate multi-level change features and semantic features with high resolution required for CD. The decoding process consists of following three steps:

Fusing and Upsampling. After feature extraction by the encoder, we obtain four features at the sizes of (h/4, w/4), (h/8, w/8), (h/16, w/16), (h/32, w/32) (h, w represents the height and width of input images). We first concatenate feature pairs hierarchically along the channel dimension to generate four change features at different levels, then decode change features by an attention module, named SCAM. Finally, we upsample low-level features to concatenate with high-level features in a bottom-up manner. As a result, we learn multi-level attention-refined change features for change detection. This step corresponds to the pseudo-code in lines 9–10 of Algorithm 1.

Our intuition of refining features is that emphasizing those features relevant to change of interest, and suppressing those features irrelevant to change. Consequently, we design a spatial-channel attention module, which is shown in Figure 2. The channel attention module (see Figure 2b) aims to capture channel-wise importance through a channel attention map, and it focuses on 'which channel' to emphasize or suppress from the concatenated feature. The channel attention module is calculated as follows:

$$M_c^{avg} = \mathrm{MLP}(\mathrm{AvgPool}(F)) \tag{1}$$

$$M_c^{max} = \mathrm{MLP}(\mathrm{MaxPool}(F)) \tag{2}$$

$$M_c(F) = \sigma\left(M_c^{avg} + M_c^{max}\right) \tag{3}$$

where $F \in \mathbb{R}^{C\times H\times W}$ denotes the input feature, C, H, W is the channel dimension, height, and width of the feature. Firstly, we perform a global average pooling and a global max pooling to obtain two vectors with the size of $C \times 1 \times 1$, denoted by M_c^{avg} and M_c^{max}. Then, a weight-shared multi-layer perception (MLP) gives weights to each channel on the two vectors. Afterward, we perform a sigmoid operator on the element-wise sum of the two vectors to obtain the channel attention map $M_c(F)$. Finally, multiply the original feature with $M_c(F)$ to obtain the channel-refined feature $F^{'}$.

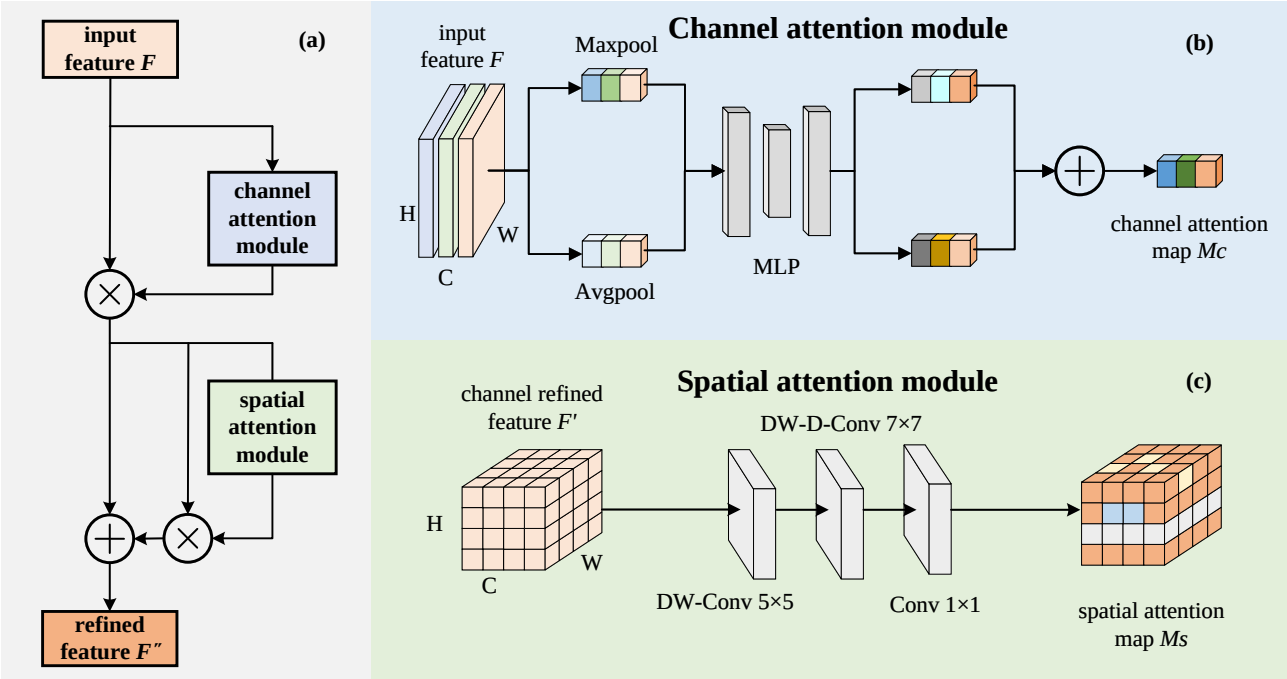

Figure 2. Architecture of the proposed SCAM: (**a**) overview of SCAM; (**b**) channel attention module, and (**c**) spatial attention module.

The spatial attention module (see Figure 2c) focuses on 'which area' to emphasize or suppress on the channel-refined feature $F^{'}$. The key step is producing an attention map which indicates the importance of each pixel. A well-known method is using large kernel convolution [51,53]. However, large kernel convolution brings a large quantity of computational cost and parameters. Inspired by the use of visual attention [54], we decompose a large kernel convolution into three convolutional layers to decrease the computational complexity and capture long-range dependence. The three convolutional layers are a spatial local convolution (depth-wise convolution, DW-Conv), a spatial long-range convolution (depth-wise dilation convolution, DW-D-Conv), and a 1×1 channel convolution. Therefore, the proposed spatial attention module can be expressed as:

$$M_s(F) = \text{Conv}_{1\times1}(\text{DW-D-Conv}_{7\times7}(\text{DW-Conv}_{5\times5}(F))) \tag{4}$$

where $F \in \mathbb{R}^{C\times H\times W}$ is the input feature, C, H, W is the channel dimension, height, and width of the feature. $\otimes$ denotes the element-wise product. The subscript of each convolutional operator represents the kernel size. $M_s(F)$ denotes the final spatial attention map. Finally, the input feature F is refined as follows:

$$F^{'} = M_c(F) \otimes F \tag{5}$$

$$F^{''} = F^{'} + M_s(F^{'}) \otimes F^{'} \tag{6}$$

We validated this module by visualizing features with Grad-CAM [55]. As shown in Figure 3, the features refined by SCAM are mostly concentrated on building areas with clearer boundaries, which indicates that our SCAM indeed calculates more accurate bitemporal features and change features.

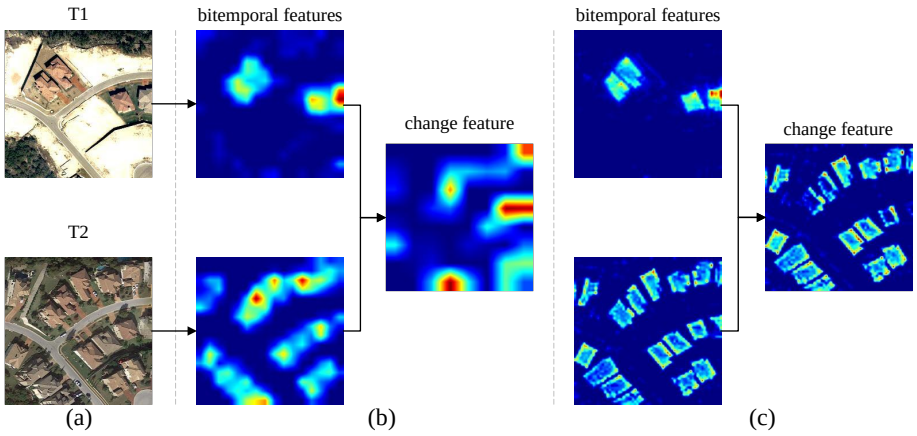

Figure 3. Feature visualization of our SCAM. Red denotes higher attention values and blue denotes lower values. (**a**) a pair of building images in the LEVIR-CD dataset; (**b**) unrefined feature maps; and (**c**) refined feature maps through multiple SCAMs.

Convolutional Forward Pass. To obtain the final change mask, most existing methods fuse extracted multi-level features by concatenation or element-wise addition. These methods are able to take advantage of multi-level features. However, simply fusing different level features may introduce misalignment to change features, which leads to noisy change borders or pseudo-changes. To overcome these cons, we propose parallel convolutional blocks to filter noise. As shown in Figure 1, the PCB is composed of gray nodes, and there are four PCBs at different levels in our network. The attention-refined multi-level change features are fed into t convolutional blocks in parallel, and the bitemporal features are processed in the same manner to extract more information of interest. Both bitemporal features and change features have four branches, channels of the four branches are 18, 36, 72, and 144, in turn, which are lower than most existing methods. The convolutional block

we used is the basic-block (shown in Figure 4) from ResNet [27]. The block consists of two 3×3 convolutional layers, and two batch normalization [56] layers followed a ReLU [57] activation layer. In this way, we further refine extracted features without downsampling to maintain its high resolution for the further decoding process. Lines 13–14 of Algorithm 1 reflect the process of the convolutional forward pass.

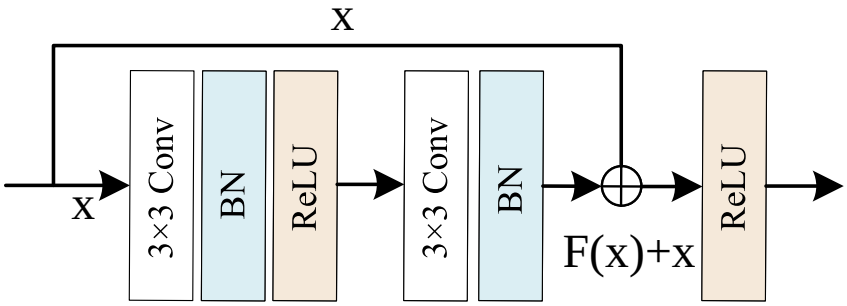

Figure 4. Structure of a basic block.

Fusing and Classification. In the wake of the high-resolution convolutional forward pass, we use hierarchically SCAMs again (see in lines 16–17 of Algorithm 1) to assemble information in the network. Then, the concatenation of multi-level change features is fed into a classifier to generate a change mask (see in line 20 of Algorithm 1). The classifier consists of two 3×3 convolutional layers, an upsampling operation, and an Argmax operation. We apply two 3×3 convolutional layers and an upsampling operation on the concatenation of multi-level change features to obtain the discriminative map $D \in \mathbb{R}^{2 \times H \times W}$. The value in the first channel of D indicates the probability that the corresponding pixel belongs to the unchange class, while the other channel indicates the probability that the corresponding pixel belongs to the change class. Finally, an Argmax operation is employed to generate the final change mask $M \in \mathbb{R}^{1 \times H \times W}$ by finding the class with the maximum value for each pixel of the discriminative map D.

The visualization of feature maps is displayed in Figure 5. It can be seen that the bitemporal features have higher values in the building areas, and there is less interference in the change feature map after refinements of PCBs and SCAMs. Hence, we can conclude that our proposed deep attentive high-resolution change decoder can extract discriminative features and model the long-range dependence for change detection effectively.

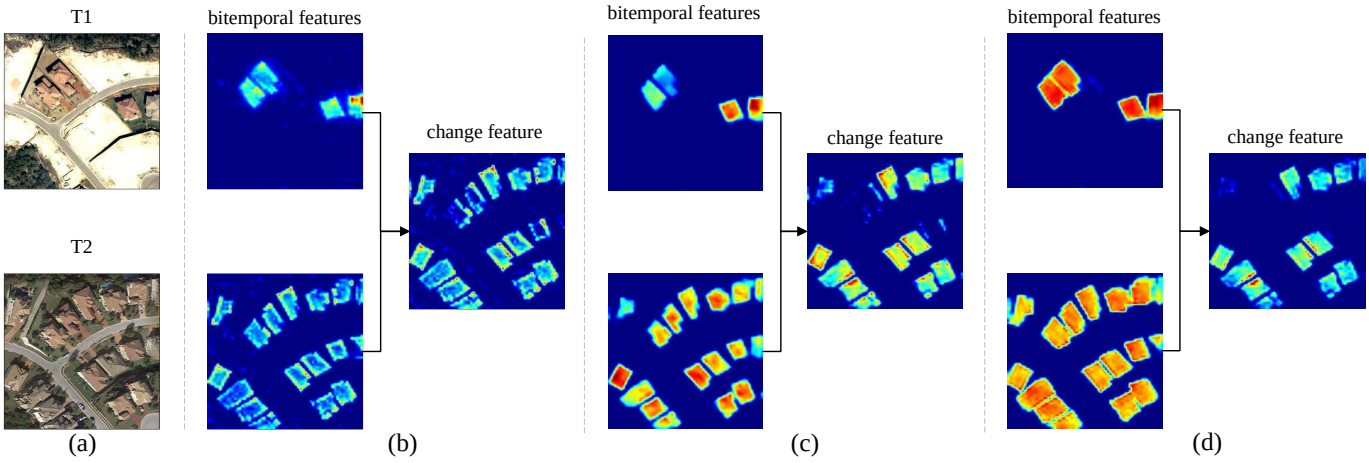

Figure 5. Feature visualization of our PCBs and SCAM. Red denotes higher attention values and blue denotes lower values. (**a**) a pair of building images in the LEVIR-CD dataset; (**b**) refined feature maps through SCAMs; (**c**) further refined feature maps through PCBs; and (**d**) feature maps that are refined by SCAMs again.

3.3. Deep Supervision and Loss Function

The batch contrastive loss (BCL) [58] and the binary cross-entropy (BCE) loss are widely used for change detection in remote sensing. The BCL is used to measure the similarity between the distance map and the ground truth, which is defined as follows:

$$L_{\mathrm{BCL}} = \sum_{i,j=0}^{N} \frac{1}{2}[(1-y_{i,j})d_{i,j}^2 + y_{i,j}\max(m-d_{i,j},0)^2] \tag{7}$$

where $d_{i,j}$ represents the Euclidean distance of the feature pair at point (i,j). $y_{i,j}$ denotes the label at point (i,j), which is 0 or 1. N is the size of the distance map. m is the margin to filter out pixel pairs with a distance lower than this value.

The BCE loss can be calculated as follows:

$$L_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i,j=0}^{N}\left[y_{i,j}\log x_{i,j} + (1-y_{i,j})\log(1-x_{i,j})\right] \tag{8}$$

where $y_{i,j}$ and $x_{i,j}$ denote the ground-truth label and the predicted probability of change class at point (i,j), respectively.

To weaken the effect of unbalanced categories, the dice coefficient loss is combined with BCE loss, which is defined as:

$$L_{\mathrm{Dice}} = 1 - \frac{2\times|X\bigcap Y|}{|X|+|Y|} \tag{9}$$

where X and Y denote the predicted change map, and the ground-truth label, $\bigcap$, represents the intersection of X and Y.

Unlike previous networks, our network has two auxiliary branches to extract bitemporal semantic features (see Figure 1) and one branch for change detection. To improve the convergence of our network by alleviating the vanishing gradient problems and learn multi-level discriminative features for intermediate layers, we introduce a deeply supervised module into the change decoder based on the auxiliary branches.

Figure 6 illustrates the calculation process of change loss, which is described in the fourth step of Section 3.1. As shown in Figure 6, the multi-level change features are concatenated to generate the change map. For the two groups of intermediate features $\{f_i^1, i=2,3,4\}$ and $\{f_i^2, i=2,3,4\}$, a Euclidean distance map d_i is calculated at the i^{th} level. We leverage a BCL to pull unchanged bitemporal features closer and push changed bitemporal features away between f_i^1 and f_i^2, and use a BCE loss to measure the similarity among probability distributions of change class and unchange class. In addition, we introduce dice coefficient loss to alleviate the class imbalance problem. The final change loss function is formulated as:

$$d_i = \sqrt{(f_i^1 - f_i^2)(f_i^1 - f_i^2)^T}, i = 2,3,4 \tag{10}$$

$$M = classifier(Concat(c_1, c_2, c_3, c_4)) \tag{11}$$

$$L = L_{\mathrm{BCE}}(M,y) + L_{\mathrm{Dice}}(M,y) + \lambda \times \sum_{i=2,3,4} L_{\mathrm{BCL}}(d_i,y) \tag{12}$$

where λ represents the weight for deep supervision. According to our experimental results in Section 4.7, λ was set to 0.1. d_i represents the Euclidean distance of the feature pair (f_i^1, f_i^2). $M \in \mathbb{R}^{1\times H\times W}$ denotes the final change mask, H, W is the height and width of the mask.

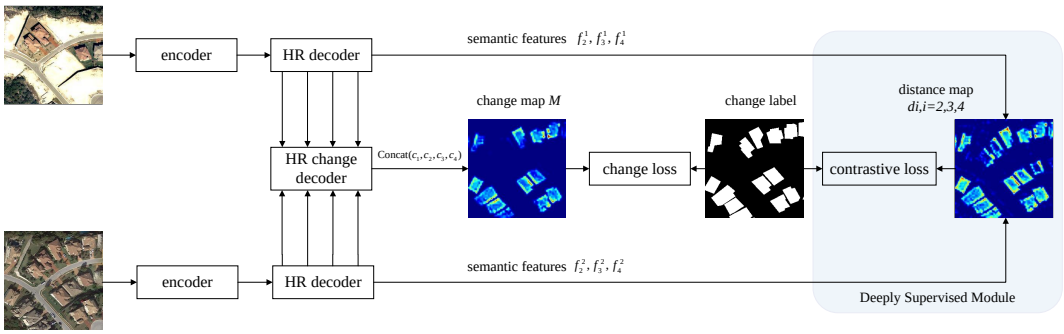

Figure 6. Deep supervision in the DSAHRNet.

4. Experiments

In this section, extensive experiments are conducted using three public change detection datasets. First, we introduce the three datasets in detail. Second, we provide details of evaluation metrics and the training process. Third, we briefly introduce eight state-of-the-art comparative methods for change detection. Finally, quantitative experiments are exhibited to demonstrate the superiority of the proposed method.

4.1. Data Sets

To fully verify the effectiveness of our method, we carried out experiments on three public datasets, which can be summarized as follows:

(1) LEVIR-CD [12] is a public large-scale building change detection dataset. It consists of 637 very high-resolution (0.5 m/pixel) image patch pairs with a size of 1024 × 1024 pixels. These bitemporal images were collected from several cities in Texas of the United States with a time span of 5–14 years. Following the official train/validation/test split, we cropped images into 256 × 256 without overlapping to obtain 7120 image patch pairs for training, 1024 pairs for validation, and 2048 pairs for testing.

(2) CDD dataset [59] is a public general change detection dataset. It consists of 11 season-varying remote sensing image pairs obtained by Google Earth. Data set was generated by cropping with image size 256 × 256 pixels. Therefore, we obtained 10,000/3000/3000 patch pairs for training/validation/testing, respectively.

(3) SYSU-CD [10] contains 20,000 pairs of 0.5 m aerial images collected from Hong Kong between the years 2007 and 2014. It includes six types of changes: groundwork before construction; suburban dilation; change of vegetation; newly built urban buildings; road expansion; and sea construction. This dataset was obtained by cropping and augmenting from original 800 image pairs. For each pair, 25 pairs of 256 × 256 size were sampled randomly. Finally, 20,000 pairs were divided into training, verification, and test sets with a ratio of 6:2:2.

4.2. Implementation Details

We implemented our DSAHRNet with the Pytorch framework, trained using an NVIDIA RTX TITAN with 24 GB of GPU memory. To improve the generalization ability of our model, each slice in the dataset was normalized. Following previous works [34], several data augmentation strategies were applied before being fed into the model. The strategies include random rotating (probability = 0.3, $\|\text{angle}\| \leq 20^\circ$), transposing (probability = 0.2), flipping (probability = 0.3), and adding Gaussian noise (probability = 0.3, mean = 0, $10 \leq \text{variance} \leq 40$).

We used an HRNet18 [52] as our encoder. The original HRNet18 has four stages, and each stage outputs a feature representation. These features will be fed into the change decoder to produce a change mask. The margin m in the BCL took a value of 2, and the weight λ for DS was set to 0.1. In the process of training, the batch size of all experiments was equal to 16 according to our experimental results in Appendix A. Following [12], we adopted an AdamW optimizer using a linear decay learning rate scheduler with an initial

learning rate of 1×10^{-4}. The learning rate was kept for the first 100 epochs and linearly decayed to 0 until trained for 200 epochs. The weight with the highest validation F1 score will be saved as the checkpoint for testing.

4.3. Evaluation Metrics

To measure the performance of the proposed method, we adopt five typical evaluation metrics: overall accuracy (OA), precision (Pre), recall (Rec), F1 score, and intersection over union (IoU). We only report these metrics with regard to the change class, and they can be expressed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{16}$$

$$\text{IoU} = \frac{\text{DetectionResult} \cap \text{GroundTruth}}{\text{DetectionResult} \cup \text{GroundTruth}} \tag{17}$$

where TP, FP, TN, and FN denote the number of true positive, false positive, true negative, and false negative, respectively. A higher F1 score represents a better overall performance of the method.

4.4. Comparative Methods

To demonstrate the superiority of DSAHRNet, the following eight SOTA change detection methods are compared:

- Fully Convolutional-Early Fusion (FC-EF) [43]: Early fusion method. The bitemporal images are concatenated along the channel dimension and processed through a UNet [26] architecture to generate a change map;
- Fully Convolutional-Siamese-Concatenation (FC-Siam-conc) [43]: Late fusion method, which is a variation of FC-EF. FC-Siam-conc extracts multi-level features with a Siamese backbone, and the same-level features are concatenated to fuse the bitemporal information;
- Fully Convolutional-Siamese-Difference (FC-Siam-diff) [43]: Late fusion method, which is similar to FC-Siam-conc. FC-Siam-diff concatenates the absolute difference of bitemporal features to fuse the bitemporal information;
- Feature Constraint Network for Change Detection (FCCDN) [34]: FCCDN proposes a dual encoder–decoder backbone for CD and introduces a self-supervised learning strategy to supervise feature learning;
- ChangeFormer [36]: Transformer-based method, which constructs a hierarchical transformer encoder to extract fine-grained features, and a multi-layer perception decoder to model long-range dependencies for CD.
- Deeply Supervised Image Fusion Network (DSIFNet) [11]: Multi-level feature fusion method, which adopts channel attention and spatial attention for more discriminative features. A multi-level deep supervision strategy is introduced to efficiently train intermediate layers and enhance the performance of the network;
- Siamese NestedUNet for Change Detection(SNUNet-CD) [33]: Multi-level feature fusion method, which is the combination of Siamese network and UNet++. To obtain high-resolution features of bitemporal images, the dense skip connection mechanism is employed between its encoder and decoder;
- Deeply Supervised Attention Metric-Based Network (DSAMNet) [10]: Metric-Based method, which integrates convolutional block attention modules [51] to extract more discriminative features.

We implemented the above CD networks using their public codes, and hyper-parameters were as consistent as possible with the original literature.

4.5. Results Evaluation

In this subsection, we validate the overall performance of DSAHRNet and comparative methods. The quantitative results of all methods on the LEVIR-CD, CDD, and SYSU-CD test sets are summarized in Table 1. As shown in Table 1, compared with other methods, the proposed method achieves better performance on the three benchmark datasets.

Table 1. Quantitative results on the three CD test sets. we mark the best score in bold.

Methods	LEVIR-CD					CDD					SYSU-CD				
	OA(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	OA(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	OA(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)
FC-EF [43]	97.26	60.66	69.35	82.89	75.52	96.88	77.55	83.78	91.25	87.36	88.88	62.89	74.69	79.92	77.22
FC-Siam-diff [43]	98.27	72.88	78.43	91.14	84.31	97.74	83.07	87.91	93.78	90.75	89.83	65.49	76.66	81.79	79.14
FC-Siam-conc [43]	98.37	73.55	80.78	89.15	84.76	97.64	82.36	87.39	93.47	90.33	89.71	65.36	76.04	**82.30**	79.05
DSAMNet [10]	98.24	73.78	75.32	**97.30**	84.91	99.06	92.52	93.52	**98.85**	96.12	89.61	64.68	76.54	80.67	78.55
DSIFNet [11]	98.88	79.98	89.70	88.07	88.87	98.65	89.00	96.15	92.29	94.18	91.17	67.43	83.80	77.54	80.55
ChangeFormer [36]	99.09	83.51	91.45	90.57	91.01	99.21	93.58	95.92	97.45	96.68	91.33	67.66	84.87	76.94	80.71
SNUNet-CD/48 [33]	99.13	84.15	91.78	91.00	91.39	99.11	92.77	95.93	96.56	96.25	91.54	69.27	82.81	80.90	81.85
FCCDN [34]	99.17	84.61	**94.28**	89.19	91.66	98.63	88.81	96.16	92.08	94.07	91.22	67.72	83.58	78.12	80.75
DSAHRNet(ours)	**99.20**	**85.38**	92.26	91.96	**92.11**	**99.27**	**94.07**	**96.38**	97.51	**96.94**	**92.48**	**71.78**	**86.18**	81.12	**83.57**

Experiment on the LEVIR-CD dataset. The quantitative results in the first sub-table of Table 1 show that DSAHRNet achieves the highest OA, IoU, and F1 scores of 99.20%, 85.38%, and 92.11%, respectively. It can be observed that the FC-EF model obtains the worst performance because of its early fusion strategy. Benefiting from encoder–decoder structure and late fusion strategy, the baselines of FC-Siam-conc and FC-Siam-diff outperform FC-EF, yielding an improvement of at least 8.79% in F1 score. Benefiting from the application of CBAM, DSAMNet achieves the highest score in recall, but its precision is limited at 75.32%, only higher than FC-EF. It can be seen in Figure 7 that the change maps predicted by DSAMNet have larger boundaries than the ground-truth label, resulting in a poor performance on the LEVIR-CD test sets. We believe that such a phenomenon is caused by concatenating misaligned features. Among these baselines, late fusion baselines (SNUNet-CD, DSIFNet, and ChangeFormer) show obvious advantages over the early fusion baseline, exhibiting significant improvements by at least 5% in terms of F1 score. FCCDN achieves an excellent balance between precision (94.28%) and recall (89.19%), through its self-supervision learning strategy and a proposed dual encoder–decoder backbone. However, FCCDN still relies on upsampling to obtain the high-resolution change map, which can be interfered by external factors around boundaries. Our DSAHRNet integrates SCAM and PCB modules to filter external interference that is irrelevant to change, and perform best on the LEVIR-CD test sets.

Six inference results on the LEVIR-CD test sets are displayed in Figure 7. Here, only the best six models are presented. We can see that our DSAHRNet performs better than comparative methods in complex scenarios. Concretely, ChangeFormer is not robust to large changed regions, e.g., Figure 7f. DSAMNet can not correctly render the shapes of changed regions; then, it produces more false positives. Though DSIFNet locates changes more precisely than DSAMNet, it generates more false negatives, e.g., Figure 7d–f. Our DSAHRNet successfully detects most changed regions with accurate boundaries, e.g., Figure 7a–d. As shown in Figure 7e, DSAHRNet can avoid more false alarms compared with other methods, and overcome the interference of other changes such as illuminations and trees. For large-scale changed regions in Figure 7f, DSAHRNet is only inferior to DSAMNet, but performs better in other scenarios, e.g., Figure 7a–e.

Experiment on the CDD dataset. The quantitative results on the CDD test sets are listed in the middle sub-table of Table 1. The proposed DSAHRNet achieves the first against other comparative methods in OA, IoU, precision, F1 score, and the third in recall, yielding a result of 96.94% in F1 score. DSAHRNet outperforms the other eight methods by 0.26–9.59% in terms of F1. It can be seen that the baseline FC-EF is still inferior to

FC-Siam-diff and FC-Siam-conc. DSAMNet integrates CBAM into the network to make features more discriminative before calculating the distance map, thus reaching the highest score of 98.95% in recall, and the fourth in F1 score (96.12%). However, DSAMNet still can not produce precise shapes of changed regions, leading to a lower score in precision and F1. DSIFNet obtains a higher precision score (96.15%) than the DSAMNet model, but other scores are lower. It can be observed in Figure 8 that change maps obtained by DSIFNet have more undetected change regions. By using the dense nodes in the decoder of UNet++, SNUNet-CD maintains its robustness on the CDD test sets, and F1 ranks third at 96.25%. However, different level features have equal attention, and there are more shallow features in the network. The shallow features will contribute more to the detection of changes, which results in limited improvement. FCCDN does not perform as well as in the LEVIR-CD dataset. We think its self-supervised learning strategy is not adapted to multi-class change detection. ChangeFormer achieves the second F1 score (96.68%) in detection performance, by using self-attention to capture long-range contexts.

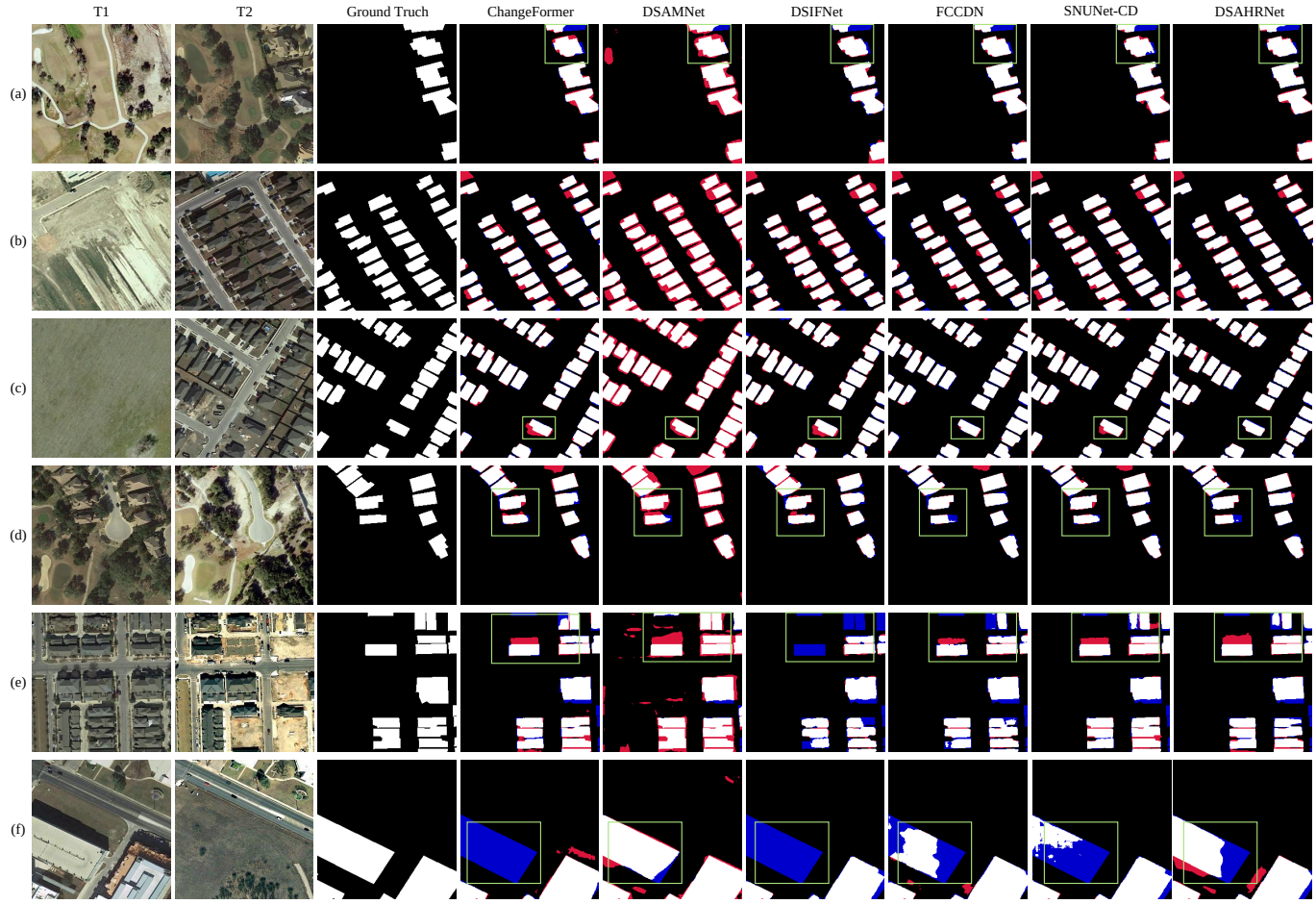

Figure 7. Six inference results of different methods on the LEVIR-CD test sets. The (**a**–**f**) indicate samples from LEVIR-CD and the change maps obtained by different methods. We highlight interesting regions with green rectangles. True positive is plotted in white, true negative is plotted in black, false positive is plotted in red, and false negative is plotted in blue.

Figure 8 provides an intuitive picture of comparative methods' performance on the CDD test sets. As shown in Figure 8, ChangeFormer, SNUNet-CD, and our DSAHRNet maintain their great performance on the SYSU-CD test sets, but our DSAHRNet generates the closest results with the ground truth. There are less omissions in the results of our DSAHRNet, and DSAHRNet performs well on multi-scale changed objects. For instance, in Figure 8a–d, most comparative methods incorrectly classify the thin and small changed

regions, such as vehicles, small buildings, and narrow roads, while DSAHRNet can distinguish them accurately. Regarding large areas of change, as seen in Figure 8e,f, DSAHRNet is able to detect large areas of changed buildings, and small areas surrounding the buildings.

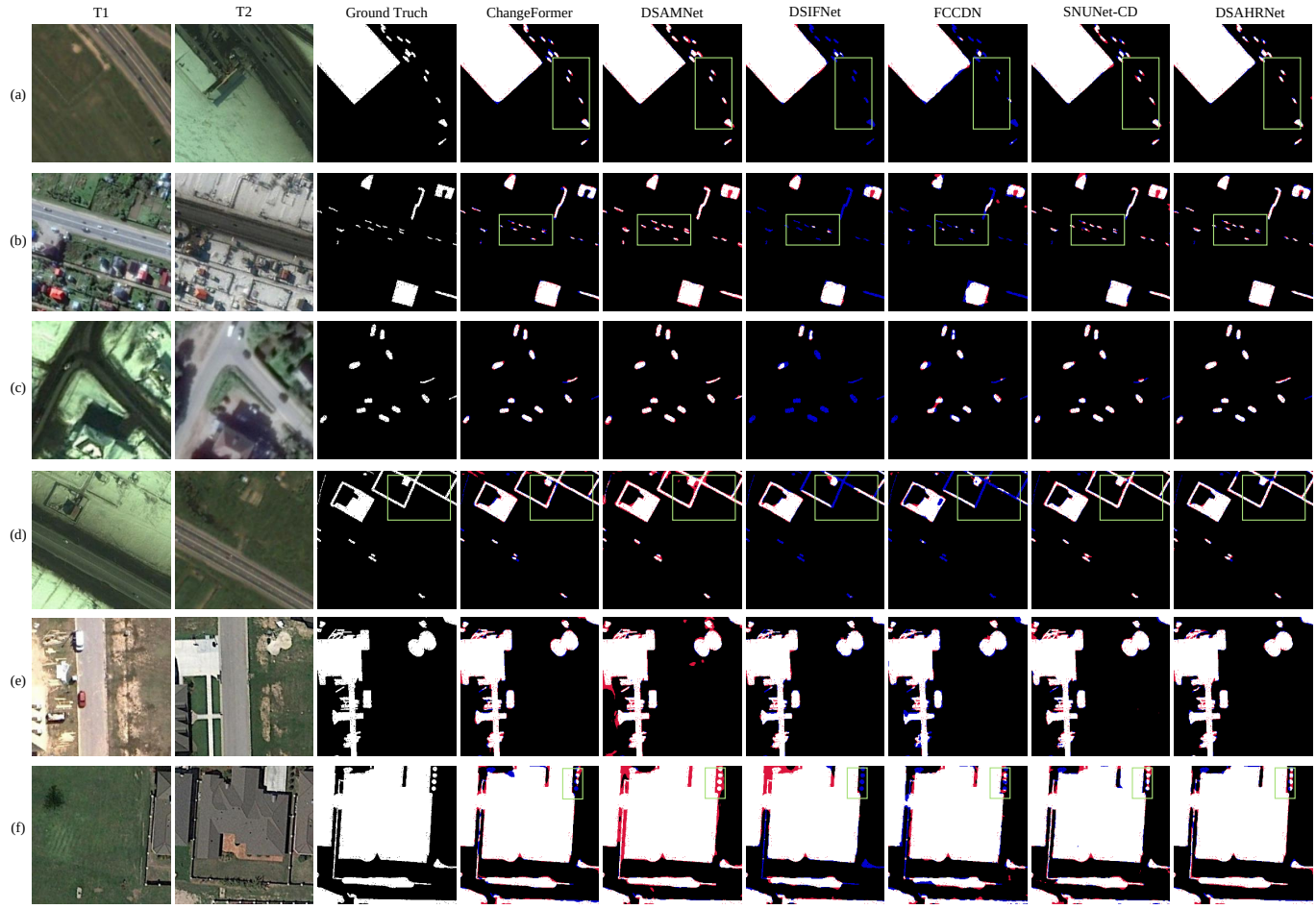

Figure 8. Six inference results of different methods on the CDD test sets. The (**a**–**f**) indicate samples from LEVIR-CD and the change maps obtained by different methods. We highlight interesting regions with green rectangles. True positive is plotted in white, true negative is plotted in black, false positive is plotted in red, and false negative is plotted in blue.

Experiment on the SYSU-CD dataset. The comparison results in the last sub-table of Table 1 show that DSAHRNet outperforms other methods with a remarkable advantage. Our DSAHRNet exhibits the highest IoU (71.78%) and F1 (83.57%) scores on the SYSU-CD test sets. Several challenging samples are illustrated in Figure 9 to verify different methods' generalization abilities. It can be seen that our DSAHRNet filters out many pseudo-changes, yielding higher scores on OA (92.48%) and Rec (81.12%). In Figure 9a–d, DSAHRNet performs well in multiple scenarios of urban, port, and forest, while other methods are sensitive to pseudo-changes from seasonal changes and illumination changes. It is noteworthy that our DSAHRNet detects a mislabeled vegetation change in Figure 9e, and presents the clearest boundary of vegetation change in Figure 9f.

4.6. Model Computation Complexity

To fairly compare the computational complexity and parameters with other works, we report F1 vs. numbers of parameters (Params) and F1 vs. multiply accumulate operations (MACs) trade-offs on the SYSU-CD test sets in Table 2. MACs can purely describe computational complexity from the mathematical perspective. To reduce computational

complexity, our DSAHRNet only generates four level features with 18, 36, 72, and 144 channels, respectively. It can be seen that our DSAHRNet obtains the highest F1 score (83.89%) with the complexity of 13.157G MACs and 14.656M parameters. visualization comparisons on F1 vs. parameters and F1 vs. MACs are displayed in Figure 10, and we can conclude that our DSAHRNet achieves the best F1 score with competitive efficiency.

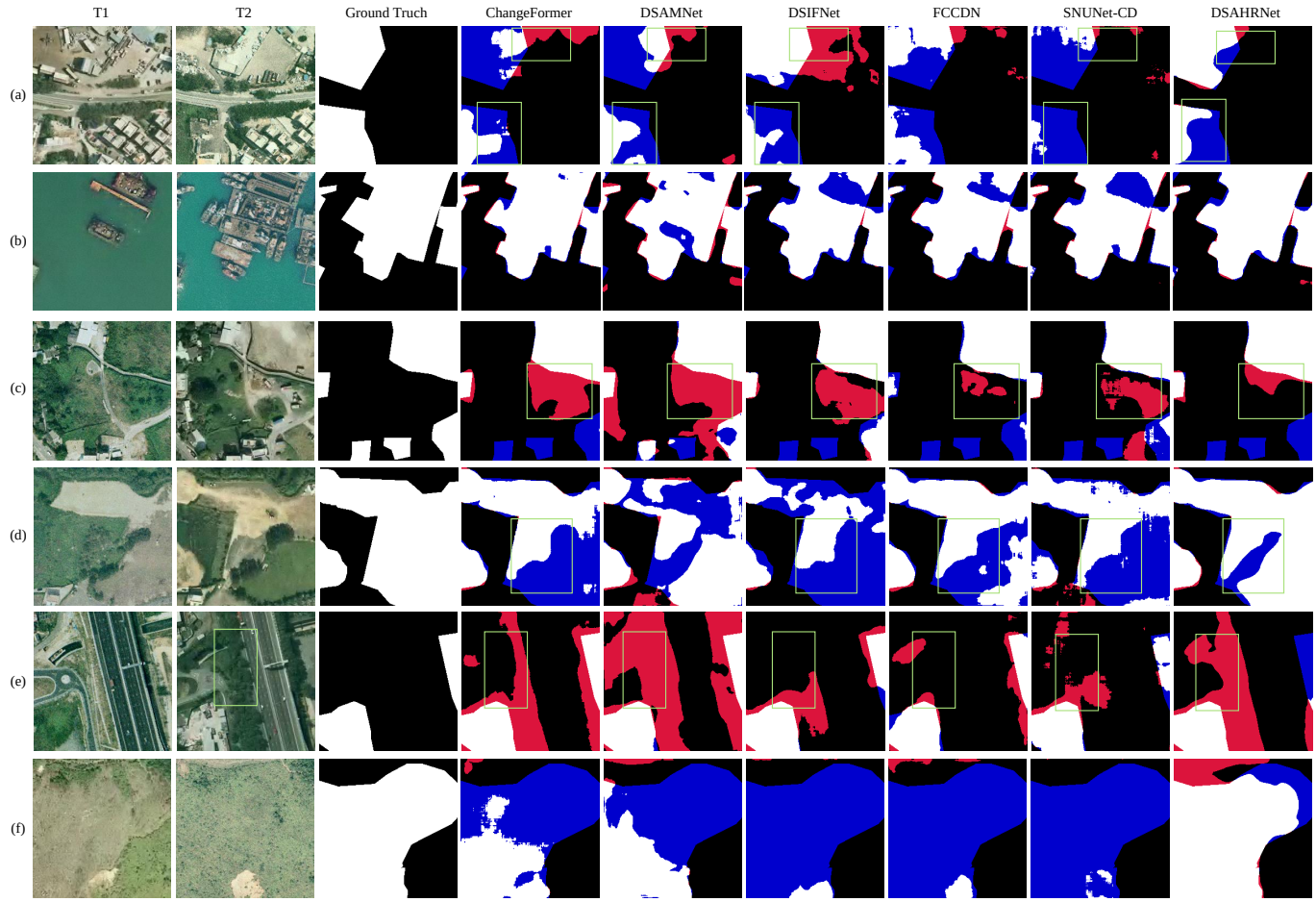

Figure 9. Six inference results of different methods on the SYSU-CD test sets. The (**a**–**f**) indicate samples from LEVIR-CD and the change maps obtained by different methods. We highlight interesting regions with green rectangles. True positive is plotted in white, true negative is plotted in black, false positive is plotted in red, and false negative is plotted in blue.

Table 2. Performance-complexity trade-off for change detection on the SYSU-CD dataset. We mark the best F1 score in bold.

Methods	MACs(G)	Params(M)	F1(%)
FC-EF [43]	3.562	1.348	77.22
DSAMNet [10]	75.29	16.951	78.55
FC-Siam-diff [43]	4.699	1.347	79.14
FC-Siam-conc [43]	5.303	1.543	79.05
DSIFNet [11]	82.264	35.728	80.55
FCCDN [34]	12.522	6.307	80.75
ChangeFormer [36]	202.829	41.005	80.71
SNUNet-CD/48 [33]	123.12	27.068	81.85
DSAHRNet(ours)	13.157	14.656	**83.57**

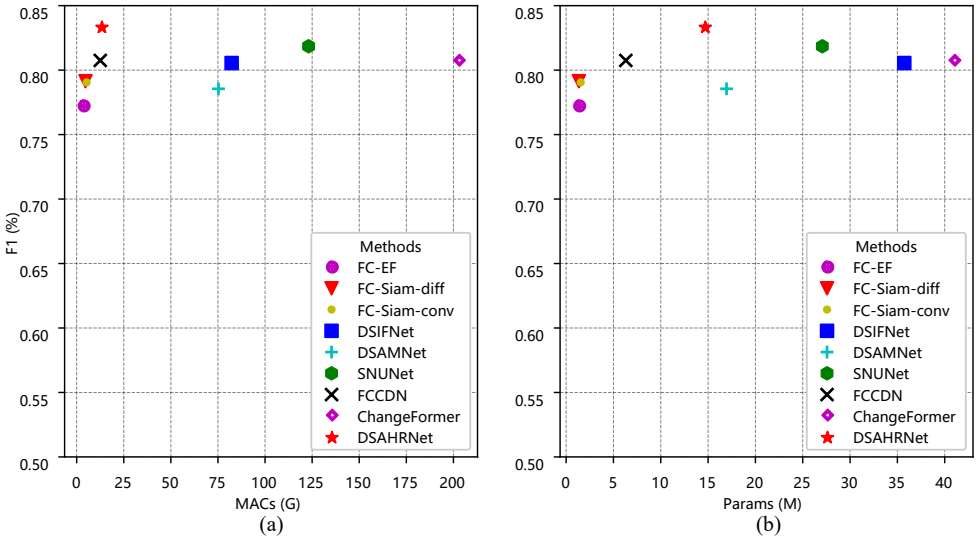

Figure 10. Illustration comparisons of performance. Here, performance is measured in F1 score. (**a**) an illustration comparison of performance and parameters, and (**b**) an illustration comparison of performance and MACs.

4.7. Ablation Study

In this subsection, we present the ablation study of DSAHRNet on the SYSU-CD dataset. Our DSAHRNet integrates SCAM, PCB, and DS modules for change detection. To validate the effectiveness of these modules, we design ablation experiments with four metrics: IoU, precision, recall, and F1 score on the SYSU-CD dataset. In the following experiments, only normalization is applied in the training stage. We denote the basic model without proposed components as 'Base', whose encoder is brought from the trained FC-Siam-conc model.

Ablation experiments of our DSAHRNet on the SYSU-CD dataset are reported in Table 3. Figure 11 illustrates the ablation result on the SYSU-CD test sets. It can be observed from Figure 11 that our proposed modules can improve semantic correctness. As shown in Table 3, the 'Base' receives the lowest F1 score of 78.43%, which is slightly lower than FC-Siam-conc of 79.05%. Results of the second line to the fourth line show the effect of a single module on the 'Base' model. First, the introduction of PCB can obviously improve F1 score of 2.2%. The result of 'Base+PCB' demonstrates that our PCB can reduce feature misalignment efficiently among bitemporal features, and extract more accurate features for change detection. Moreover, by integrating SCAM into the 'Base' model, 'Base+SCAM' achieves an improvement in F1 score of 1.9%, which shows that our SCAM can decode change features from bitemporal features within the spatial-temporal domain. Finally, our DS module can optimize the training process of DSAHRNet by making the feature pairs more distinguishable in the embedding space, 'Base+DS' achieves an improvement in F1 score of 1.21%. We set the hyper-parameter λ between (0, 0.1) to explore the sensitivity of the proposed DS module on the SYSU-CD dataset. As shown in Table 4, the F1 score reaches the highest F1 score when λ is set to 0.1. Other combinations of the proposed modules are tried, and all the combinations can achieve distinct accuracy improvements. The combination of 'Base+PCB+SCAM+DS' achieves the highest score in F1 (81.89%) among a series of 'Base' models. It is noteworthy that 'Base+PCB+SCAM+DS' is slightly higher than the second model, SNUNet-CD, in the F1 score of 0.04%. Row 8 to row 10 of Table 3 present the quantitative results with different encoders, and we obtain the best model for change detection when we replace the 'Base' encoder with HRNet18.

Table 3. Ablation study of our DSAHRNet on the SYSU-CD dataset. Values in bold font are the best. We present the performance shift with ±.

Methods	IoU(%)	Pre(%)	Rec(%)	F1(%)
Base	64.51	78.60	78.26	78.43 ± 0.03
Base+SCAM	67.12	79.96	80.70	80.33 ± 0.06
Base+PCB	67.56	82.15	79.19	80.64 ± 0.08
Base+DS	66.17	80.18	79.10	79.64 ± 0.02
Base+PCB+DS	68.40	81.93	80.56	81.24 ± 0.04
Base+PCB+SCAM	68.55	82.46	80.25	81.34 ± 0.05
Base+DS+SCAM	67.22	81.73	79.11	80.40 ± 0.11
Base+PCB+SCAM+DS	69.33	82.57	81.22	81.89 ± 0.12
van_tiny [54]+PCB+SCAM+DS	68.91	82.19	81.00	81.59 ± 0.03
HRNet18 [52]+PCB+SCAM+DS	**71.47**	**85.16**	**81.64**	**83.36** ± 0.05

Table 4. Effect of hyperparameter λ on the performance of the proposed DSAHRNet for the SYSU-CD dataset.

Methods	λ	IoU(%)	Pre(%)	Rec(%)	F1(%)
Base	0	64.51	78.60	78.26	78.43
Base+DS	0.001	66.12	80.24	78.98	79.60
Base+DS	0.01	63.14	79.50	75.42	77.41
Base+DS	0.05	65.47	79.96	79.96	79.13
Base+DS	0.1	66.17	80.18	79.10	79.64

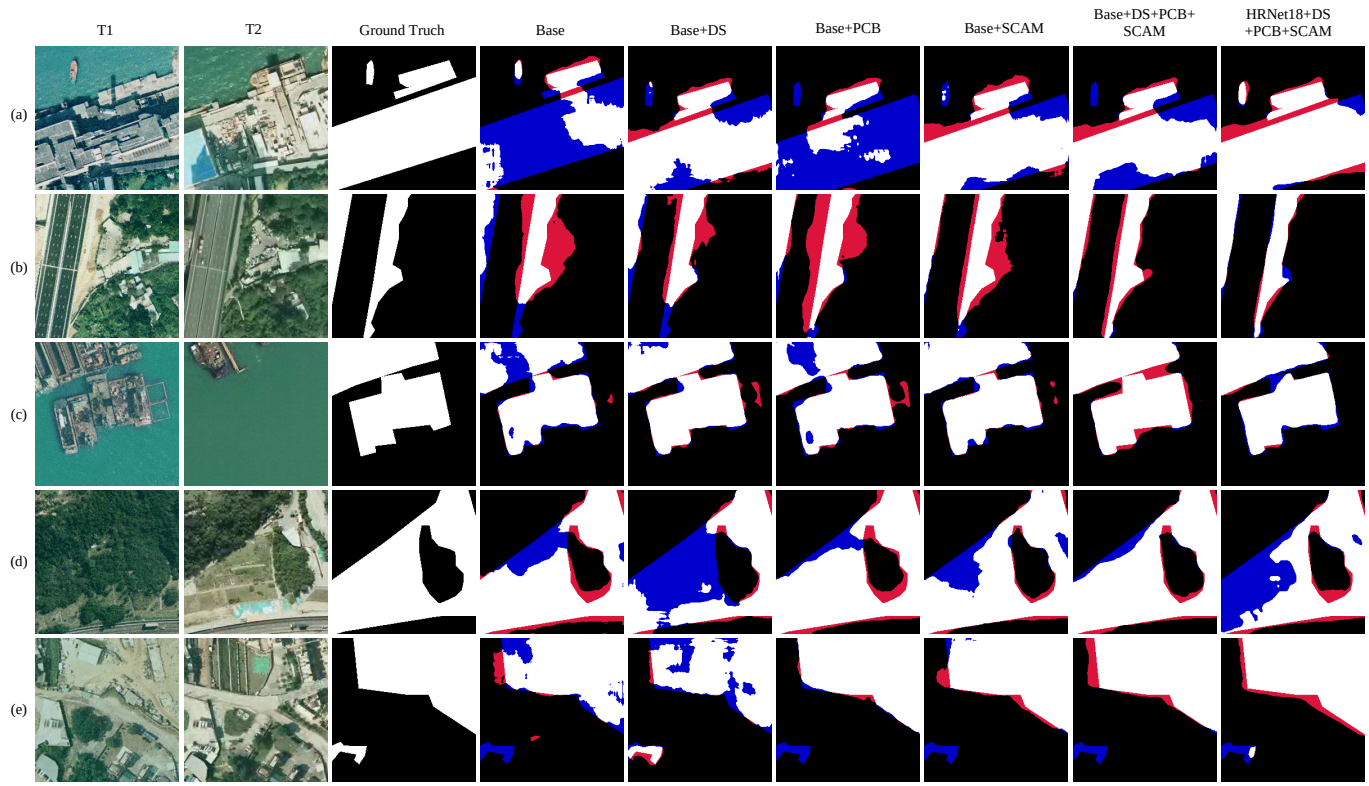

Figure 11. Ablation results of the proposed method on the SYSU-CD test sets. The (**a**–**e**) indicate samples from LEVIR-CD and the change maps obtained by different methods. True positive is plotted in white, true negative is plotted in black, false positive is plotted in red, and false negative is plotted in blue.

5. Discussion

In this paper, extensive experiments are implemented on three benchmark datasets, LEVIR-CD, CDD, and SYSU-CD. Through quantitative comparison, our DSAHRNet is superior to other methods in F1-score, and achieves a better trade-off between performance and complexity than other methods. Our DSAHRNet has good adaptability to multi-scale change areas due to the introductions of DS, PCB, and SCAM; these modules enable the network to learn discriminative features. Our method can be considered as a combination of the metric-based and classification-based method due to the introduction of the DS module. It is a metric-based method when we only calculate a distance map from bitemporal features to generate a change map, and it is a classification-based method when we only calculate probability distributions of change class and unchange class. At the same time, it should be pointed out that the network also has some potential limitations. First, we choose a tiny model, HRNet18, as our feature encoder for a better trade-off between performance and complexity. When a deeper encoder, e.g., ResNet50, HRNet48, is employed, the computational complexity of the DSAHRNet will increase significantly because there are a lot of convolutional layers in the change decoder. Second, DSAHRNet needs sufficient and effective training data for training. Table 5 reports the adaptability of the proposed method with 10% and 50% labeled images. The results show that the proposed method has different degrees of performance degradation to partially labeled datasets, and there is still a large room for improvement. Therefore, a future direction may be reducing the model complexity in the change decoder by model compression techniques. Moreover, we will conduct further research on semi-supervised or unsupervised learning to facilitate the practical application of change detection in remote sensing.

Table 5. The adaptability of the proposed model to different numbers of labeled images.

Labeled Images	LEVIR-CD					CDD					SYSU-CD				
	OA(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	OA(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	OA(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)
10%	97.83	67.30	74.27	87.77	80.46	95.31	66.59	80.74	79.17	79.95	89.07	63.10	75.56	79.29	77.38
50%	98.68	78.48	82.49	94.17	87.94	97.90	83.82	90.12	92.30	91.20	91.44	69.45	81.46	82.48	81.97
100%	99.20	85.38	92.26	91.96	92.11	99.27	94.07	96.38	97.51	96.94	92.48	71.78	86.18	81.12	83.57

6. Conclusions

In this paper, a novel deep learning network (DSAHRNet) is proposed for change detection in remote sensing images. To extract change features precisely, our DSAHRNet is constructed on the encoder–decoder structure with the proposed SCAM and PCB. The SCAM exploits attention maps to enhance the process of feature fusion. The PCB refines extracted features to make them more discriminative. Moreover, a DS strategy is introduced to optimize feature learning. Compared with existing SOTA methods, our DSAHRNet achieved the best performance on three benchmark datasets, and a better trade-off between complexity and performance, which indicates that the proposed method is robust and effective for change detection.

Author Contributions: Conceptualization, J.W.; methodology, J.W.; software, J.W., Z.Z. and C.X.; formal analysis, J.W.; investigation, J.W. and C.X.; resources, J.W.; data curation, J.W. and C.X.; writing—original draft preparation, J.W.; writing—review and editing, J.W., C.X., Z.Z. and Y.Z.; visualization, J.W. and C.X.; supervision, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The code for this study will available at https://github.com/Githubwujinming/DSAHRNet (accessed on 13 October 2022). The LEVIR-CD, CDD, and SYSU-CD datasets are openly available at https://justchenhao.github.io/LEVIR/ (accessed on 13 October 2022), https://drive.google.com/file/d/1GX656JqqOyBi_Ef0w65kDGVto-nHrNs9 (accessed on 13 October 2022) and https://mail2sysueducn-my.sharepoint.com/:f:/g/personal/liumx23_mail2_sysu_edu_cn/Emgc0jtEcshAnRkgq1ZTE9AB-kfXzSEzU_PAQ-5YF8Neaw?e=IhVeeZ (accessed on 13 October 2022)

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

Abbreviations

The following abbreviations are used in this manuscript:

DSAHRNet	Deeply Supervised Attentive High-Resolution Network
FC-EF	Fully Convolutional-Early Fusion
FC-Siam-Conc	Fully Convolutional-Siamese-Concatenation
FC-Siam-diff	Fully Convolutional-Siamese-Difference
FCCDN	Feature Constraint Network for Change Detection
DSIFNet	Deeply Supervised Image Fusion Network
SNUNet-CD	Siamese NestedUNet for Change Detection
DSAMNet	Deeply Supervised Attention Metric-Based Network
CBAM	Convolutional Block Attention Module
GT	Ground Truth
TP	True Positive
TN	True Negative
FP	False Positive
SCAM	Spatial-Channel Attention Module
PCB	Parallel Convolution Block
RS	Remote Sensing
CD	Change Detection
DS	Deep Supervision
BCE	Binary Cross-Entropy
BCL	Batch Contrastive Loss
CNN	Convolutional Neural Network

Appendix A. Qualitative Results of Our Model with Different Hyperparameters on the SYSU-CD Dataset

In this section, we provide additional qualitative results of our model with different hyperparameters on the SYSU-CD dataset. First, we adapt the batch size from 4 to 32 for the SYSU-CD dataset with the learning rate of 1×10^{-4}, and only normalization is applied in the training stage. As shown in Figure A1, our model is stable to different batch sizes. Next, we adapt the learning rate from 1×10^{-6} to 1×10^{-3} with the batch size of 16, Figure A2 shows the performance of our network on the SYSU-CD dataset with different learning rates, and the F1 score reaches the peak when the learning rate approaches 1×10^{-4}. Considering the training speed and experimental consistency, the batch size is set to 16, and the learning rate is set to 1×10^{-4} in our experiments.

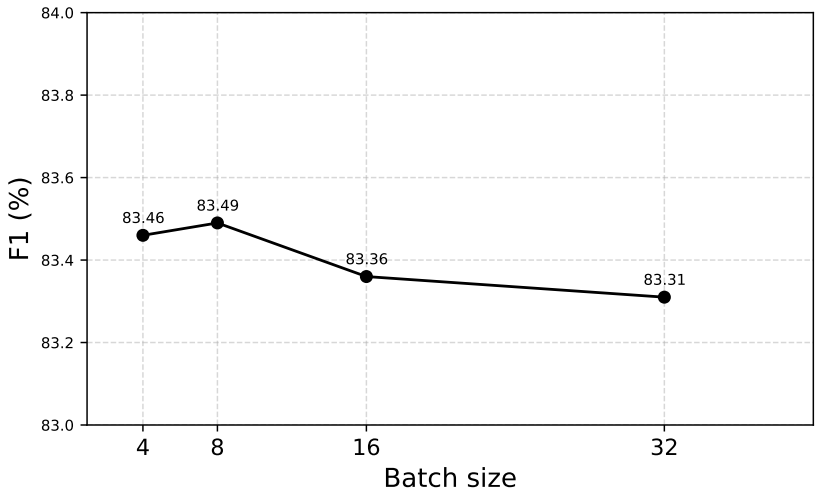

Figure A1. Qualitative results of our model with different batch sizes on the SYSU-CD dataset.

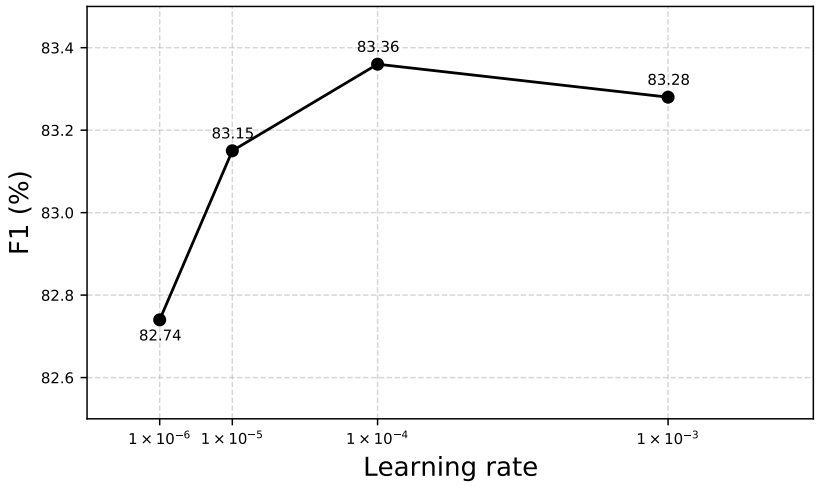

Figure A2. Effect of learning rate on the performance of the proposed DSAHRNet for the SYSU-CD dataset.

References

1. Singh, A. Review article digital change detection techniques using remotely-sensed data. *Int. J. Remote Sens.* **1989**, *10*, 989–1003. [CrossRef]
2. Marin, C.; Bovolo, F.; Bruzzone, L. Building change detection in multitemporal very high resolution SAR images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2664–2682. [CrossRef]
3. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 811–815. [CrossRef]
4. Ding, Y.; Zhang, Z.; Zhao, X.; Hong, D.; Li, W.; Cai, W.; Zhan, Y. AF2GNN: Graph convolution with adaptive filters and aggregator fusion for hyperspectral image classification. *Inf. Sci.* **2022**, *602*, 201–219. [CrossRef]
5. Sublime, J.; Kalinicheva, E. Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the Tohoku tsunami. *Remote Sens.* **2019**, *11*, 1123. [CrossRef]
6. Yao, D.; Zhi-li, Z.; Xiao-feng, Z.; Wei, C.; Fang, H.; Yao-ming, C.; Cai, W.W. Deep hybrid: Multi-graph neural network collaboration for hyperspectral image classification. *Def. Technol.* 2022. *in press*. [CrossRef]
7. Ding, Y.; Zhao, X.; Zhang, Z.; Cai, W.; Yang, N. Graph sample and aggregate-attention network for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5504205. . [CrossRef]
8. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sens. Environ.* **2021**, *265*, 112636. [CrossRef]
9. Mahdavi, S.; Salehi, B.; Huang, W.; Amani, M.; Brisco, B. A PolSAR change detection index based on neighborhood information for flood mapping. *Remote Sens.* **2019**, *11*, 1854. [CrossRef]
10. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
11. 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]
12. 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]
13. Benedek, C.; Szirányi, T. Change detection in optical aerial images by a multilayer conditional mixed Markov model. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3416–3430. [CrossRef]
14. Daudt, R.C.; Le Saux, B.; Boulch, A.; Gousseau, Y. Urban change detection for multispectral earth observation using convolutional neural networks. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2115–2118.
15. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A general end-to-end 2D CNN framework for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 3–13. [CrossRef]
16. Johnson, R.D.; Kasischke, E. Change vector analysis: A technique for the multispectral monitoring of land cover and condition. *Int. J. Remote Sens.* **1998**, *19*, 411–426. [CrossRef]
17. Zhang, J.; Zhang, Y. Remote sensing research issues of the national land use change program of China. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 461–472. [CrossRef]
18. Wu, C.; Du, B.; Zhang, L. Slow feature analysis for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 2858–2874. [CrossRef]

19. 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]
20. 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]
21. Gapper, J.J.; El-Askary, H.; Linstead, E.; Piechota, T. Coral Reef change Detection in Remote Pacific islands using support vector machine classifiers. *Remote Sens.* **2019**, *11*, 1525. [CrossRef]
22. Zong, K.; Sowmya, A.; Trinder, J. Building change detection from remotely sensed images based on spatial domain analysis and Markov random field. *J. Appl. Remote Sens.* **2019**, *13*, 024514. [CrossRef]
23. Im, J.; Jensen, J.R. A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sens. Environ.* **2005**, *99*, 326–340. [CrossRef]
24. Wessels, K.J.; Van den Bergh, F.; Roy, D.P.; Salmon, B.P.; Steenkamp, K.C.; MacAlister, B.; Swanepoel, D.; Jewitt, D. Rapid land cover map updates using change detection and robust random forest classifiers. *Remote Sens.* **2016**, *8*, 888. [CrossRef]
25. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
26. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
27. 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.
28. Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; Ding, H.; Huang, X. SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5891–5906. [CrossRef]
29. Bandara, W.G.C.; Patel, V.M. Revisiting Consistency Regularization for Semi-supervised Change Detection in Remote Sensing Images. *arXiv* **2022**, arXiv:2204.08454.
30. Chen, Y.; Bruzzone, L. Self-supervised Remote Sensing Images Change Detection at Pixel-level. *arXiv* **2021**, arXiv:2105.08501.
31. Zheng, Z.; Ma, A.; Zhang, L.; Zhong, Y. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15193–15202.
32. Caye Daudt, R.; Le Saux, B.; Boulch, A.; Gousseau, Y. Guided anisotropic diffusion and iterative learning for weakly supervised change detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
33. 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]
34. 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]
35. Zhang, M.; Xu, G.; Chen, K.; Yan, M.; Sun, X. Triplet-based semantic relation learning for aerial remote sensing image change detection. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 266–270. [CrossRef]
36. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. *arXiv* **2022**, arXiv:2201.01293.
37. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
38. Jiang, H.; Hu, X.; Li, K.; Zhang, J.; Gong, J.; Zhang, M. PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection. *Remote Sens.* **2020**, *12*, 484. [CrossRef]
39. 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]
40. Wang, M.; Tan, K.; Jia, X.; Wang, X.; Chen, Y. A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images. *Remote Sens.* **2020**, *12*, 205. [CrossRef]
41. 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]
42. Xu, J.; Luo, C.; Chen, X.; Wei, S.; Luo, Y. Remote sensing change detection based on multidirectional adaptive feature fusion and perceptual similarity. *Remote Sens.* **2021**, *13*, 3053. [CrossRef]
43. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
44. Liu, R.; Jiang, D.; Zhang, L.; Zhang, Z. Deep depthwise separable convolutional network for change detection in optical aerial images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1109–1118. [CrossRef]
45. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
46. 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]

47. Stoyanov, D.; Taylor, Z.; Carneiro, G.; Syeda-Mahmood, T.; Martel, A.; Maier-Hein, L.; Tavares, J.M.R.; Bradley, A.; Papa, J.P.; Belagiannis, V.; et al. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11045.
48. Zhang, Y.; Fu, L.; Li, Y.; Zhang, Y. Hdfnet: Hierarchical dynamic fusion network for change detection in optical aerial images. *Remote Sens.* **2021**, *13*, 1440. [CrossRef]
49. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7296–7307. [CrossRef]
50. Diakogiannis, F.I.; Waldner, F.; Caccetta, P. Looking for change? Roll the dice and demand attention. *Remote Sens.* **2021**, *13*, 3707. [CrossRef]
51. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
52. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
53. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
54. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *arXiv* **2022**, arXiv:2202.09741.
55. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
56. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 448–456.
57. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
58. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
59. Lebedev, M.; Vizilter, Y.V.; Vygolov, O.; Knyaz, V.; Rubis, A.Y. Change detection in remote sensing images using conditional adversarial networks. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, Riva del Garda, Italy, 4–7 June 2018; Volume 42.