HANet: A Hierarchical Attention Network for Change Detection With Bitemporal Very-High-Resolution Remote Sensing Images

Chengxi Han, *Student Member, IEEE*, Chen Wu, *Member, IEEE*, Haonan Guo, *Student Member, IEEE*, Meiqi Hu, *Graduate Student Member, IEEE*, and Hongruixuan Chen, *Student Member, IEEE*

***Abstract*—Benefiting from the developments in deep learning technology, deep-learning-based algorithms employing automatic feature extraction have achieved remarkable performance on the change detection (CD) task. However, the performance of existing deep-learning-based CD methods is hindered by the imbalance between changed and unchanged pixels. To tackle this problem, a progressive foreground-balanced sampling strategy on the basis of not adding change information is proposed in this article to help the model accurately learn the features of the changed pixels during the early training process and thereby improve detection performance. Furthermore, we design a discriminative Siamese network, hierarchical attention network (HANet), which can integrate multiscale features and refine detailed features. The main part of HANet is the HAN module, which is a lightweight and effective self-attention mechanism. Extensive experiments and ablation studies on two CD datasets with extremely unbalanced labels validate the effectiveness and efficiency of the proposed method.**

***Index Terms*—Attention mechanism, change detection (CD), convolutional Siamese network, remote sensing (RS) image, very-high-resolution (VHR).**

I. Introduction

CHANGE detection (CD) is the process of identifying differences in the state of an object or phenomenon by observing it at different times [1]. The goal of binary CD is to assign binary labels (i.e., change or no change) to every pixel in a region [1]. Very-high-resolution (VHR) remote sensing (RS) imagery CD is one of the fundamental topics in the field of RS image interpretation, and has a wide range of applications, including land use land cover analysis [2], urban extension studies [3], environmental monitoring [4], and disaster assessment [5]. Notably, the feature extraction of temporal RS images will be negatively influenced due to differences in the light illumination, contrast, quality, resolution, and noise of RS images in the same region at different times, which makes RS image CD a challenging field of research.

CD is one of the hot topics in the field of RS and can be broadly divided into traditional methods and deep learning methods. Researchers have devised a large number of traditional and deep learning methods of CD. Among these, the traditional methods depend primarily on original image information and handcrafted features. Bruzzone and Prieto [6] propose the change vector analysis technique to calculate the intensity and direction of two or more types of change. Nielsen et al. [7] propose the multivariate alteration detection method to maximize the variance of the transformed variable; this approach is insensitive to affine transformations. In [8], principal component analysis is proposed to convert a differential or stacked image into a new feature space, making this approach a feature transformation method. Wu et al. [9] propose the use of slow feature analysis to extract the most invariant component from a multitemporal RS image and convert the original image into a new feature space, in which the changed pixels are highlighted, whereas the unchanged pixels are suppressed.

Subsequently, the method of postclassification comparison was developed. Through sample selection and the feature classification of two RS images, the detailed change information of features can be obtained by comparing the classification map [10]. However, the classification comparison method often requires a large amount of training data; moreover, the detection accuracy depends entirely on the initial classification accuracy.

Although these traditional CD methods have achieved good detection results in their respective application scenarios, due to the limitations of these traditional manually designed feature models, they only use the spectral information of multitemporal image data. Thus, traditional methods are mainly utilized on medium- and low-resolution RS images. In comparison, VHR images contain many more spatial details; as a result, traditional methods will lead to internal fragmentation phenomena appearing in the changed area and salt and pepper noise in the unchanged area.

The generalization and accuracy of the traditional methods for big-data CD tend to be limited since they depend primarily on original image information and handcrafted features. For

Manuscript received 19 January 2023; revised 22 March 2023; accepted 26 March 2023. Date of publication 5 April 2023; date of current version 25 April 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903300 and Grant 2022YFB3903405, and in part by the National Natural Science Foundation of China under Grant T2122014, Grant 61971317, Grant 62225113, and Grant 42230108. *(Corresponding author: Chen Wu.)*

Chengxi Han, Chen Wu, Haonan Guo, and Meiqi Hu are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: chengxihan@whu.edu.cn; chen.wu@whu.edu.cn; guohnwhu@163.com; meiqi.hu@whu.edu.cn).

Hongruixuan Chen is with the Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan (e-mail: qschrx@gmail.com).

Our model is available at https://github.com/ChengxiHAN/HANet-CD.

Digital Object Identifier 10.1109/JSTARS.2023.3264802

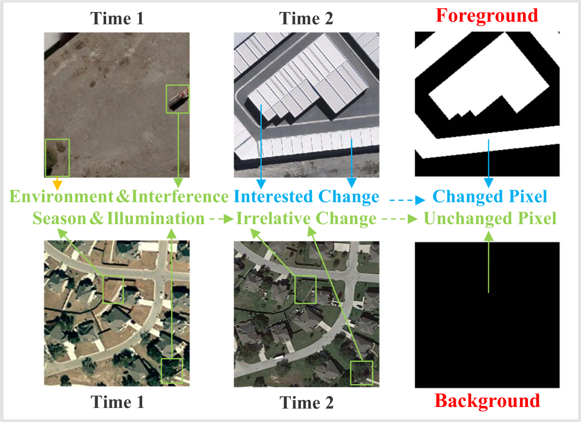

Fig. 1. Main data categories for VHR image CD include foreground and background images. The foreground image is an interesting change containing a region of interest, such as a building. The background image is an irrelevant change containing seasonal, illumination, environmental, and other interference changes.

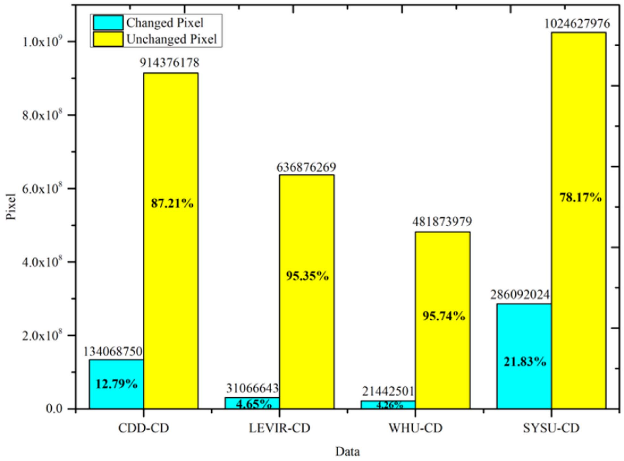

Fig. 2. Statistics related to the number of changed and unchanged pixels (CDD-CD, LEVIR-CD, WHU-CD, and SYSU-CD).

this reason, researchers have worked to develop automatic CD methods based on deep learning, which can achieve better performance.

Deep learning for CD is one of the current hottest topics in the field of RS, as it enables the extraction of representative deep features. In [11], three fully convolutional neural network (CNN) architectures (FC-EF, FC-Siam-conc, and FC-Siam-diff) are proposed. Two of these are based on Siamese networks; this represents the first time that a skip connection has been added on the basis of a Siamese network. Moreover, the attention mechanism [12] enables the model to focus on, and to fully learn and absorb, important information, which is then introduced to the CD; some works employing this approach include STANet [13], SNUNet [14], and MSPSNet [15]. Over the years, the transformer mechanism [16] has been very popular in the field of natural language processing and was subsequently introduced into the computer vision and CD contexts. Examples include BIT [17], Change Former [18], and RSP-BIT [19]. Notably, while these deep learning methods all achieve good results in their own scenarios, they do not focus on the problem of sample imbalance.

As shown in Fig. 1, the goal of CD is to find the changed pixels (white) and unchanged pixels (black). In general, there are two categories used in VHR image CD, namely foreground and background images. The foreground is a relevant change containing a region of interest such as a building. The background is an irrelevant change due to seasonal, illumination, environmental, and other interference changes.

At the same time, as shown in Fig. 2, we compile statistics related to the number of changed and unchanged pixels (CDD-CD [20], LEVIR-CD [13], WHU-CD [21], and SYSU-CD [22]). From Fig. 2, it can be observed that the percentage of changed pixels is extremely low; we refer to this as an *extremely unbalanced binary CD*.

However, most existing methods tend to ignore the imbalance problem and learn the features from the data directly, which leads to the model cannot sufficiently learn the features of the data. A few existing methods have attempted to solve the imbalance problem by designing new loss functions (e.g., dice loss and weighted cross-entropy loss). In addition, some attention-based methods struggle to relate to long-range concepts in space-time, whereas the computational complexity of some self-attention (SA) based methods is unacceptably high. Thus, it is necessary to develop more efficient and lightweight attention mechanisms. And also some methods synthesize CD samples like CDNet+IAug [23] and transfer a pretrained model like SaDL [24] to make it more robust, In fact, it is the addition of changing pixels to improve the effectiveness of the model.

To better solve these problems, we propose a set of progressive foreground-balanced sampling (PFBS) approaches on the basis of not adding change information to help the model accurately learn the features of the foreground image during the early stages of the training process, and design a more discriminative Siamese network, hierarchical attention network (HANet), to integrate multiscale features and refine detailed features. The main contributions of this work can be summarized as follows.

1) An original PFBS strategy on the basis of not adding change information is put forward to deal with the data-imbalance challenge of binary CD without additional computation cost. The progressive policy first studies the minority foreground samples in a centralized manner, empowering the network to learn the most significant characteristic of change features.
2) A discriminative Siamese HANet is tailored to integrate multiscale features and refine detailed spatial and temporal change features, where a lightweight and effective HAN module is capable of capturing long-term dependencies separately from the column and row dimensions.
3) Extensive experiments and ablation studies on two extremely unbalanced binary CD datasets validate the effectiveness and efficiency of the proposed method. We open-source the code and hope to contribute to the field of CD research.

The rest of this article is organized as follows. Section II describes the previous works related to deep-learning-based methods and the attention mechanism in the CD context. In Section III, the structure of the proposed HANet is illustrated in detail. Section IV reports the experimental results and ablation study. Finally, Section V concludes the article.

II. Related Work

In this section, we briefly introduce the deep-learning-based method and the attention mechanism of CD.

A. *Deep-Learning-Based CD Methods*

Deep CNNs can extract hierarchical features. Among them, high-level features contain abstract semantic information, whereas low-level features contain rich spatial details. Therefore, high-level semantic information is crucial to understanding scenes with complex backgrounds. This has led to the rise of RS interpretation methods based on deep learning, such as object detection [25], [26], classification [27], semantic segmentation [28], [29], and CD [30], [31].

Performing CD on RS images requires pixel-level prediction, and the fully CNN [32] is mainly used for intensive prediction tasks. Therefore, CD methods based on deep learning mainly use structures similar to the fully convolutional network, then generate difference maps or change vectors by comparing features of different depths. In [33], a deep Siamese CNN was used to capture similar features of bitemporal RS images, after which the K-nearest neighbor method was used to cluster similar pixels in the feature map to obtain the regions of variation.

These methods are a combination of deep neural networks and machine learning algorithms. Notably, although these methods have achieved better results, the whole process cannot achieve end-to-end training. In [34], bitemporal RS images were combined into one image as input to the modified U-Net [35] network structure. In [36], the improved semantic segmentation network UNet++ [37] was introduced for application to the CD task. Moreover, in order to obtain global information and fine boundary information, deeply supervised image fusion strategies are used in the encoder structure, which significantly improves the CD of VHR RS images. In [11], a Siamese fully convolutional network structure was proposed to extract the feature information of image pairs by using two identical network structures in the same manner as shared weights. However, these methods do not focus on the extremely unbalanced samples in the CD task and the rules of learning the features of the deep learning model.

B. *Attention Mechanism*

The attention mechanism is an important module in machine learning and is widely used in various types of machine learning tasks, including natural language processing [12], image recognition [38], and speech recognition [39]. The attention mechanism helps the model to assign different weights to each part of the input and extract more critical and important information. It can ensure that the model makes a more accurate judgment, and at the same time, does not introduce any additional calculation and storage overhead. Similarly, the attention mechanism has been introduced into each of the RS CD tasks listed above. In [40], spatial attention modules and channel attention modules (CAM) were used to improve the boundary integrity and semantic consistency of the change feature map. In [13], the performance of CD is improved by exploring the relationship between pixels in different channels and space, and the SA mechanism is used to calculate the weights of two pixels in different bitemporal images to generate discriminative features. In [41], a model based on super-resolution and stacking attention is proposed to improve the CD performance. In [42], a method based on differential image guidance and an attention model is proposed to describe the interval correlation of low-level and high-level features. Although these methods can improve the details of semantic information in CD tasks, there is still considerable room for improvement and optimization.

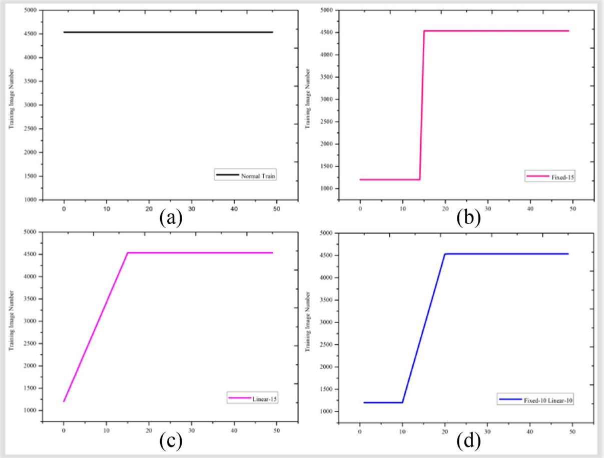

Fig. 3. Visualization of the proposed PFBS process.

III. Proposed HANet

In this section, we first introduce the motivation behind the proposed method and then present the model details. Figs. 3 and 4 show the proposed PFBS approach and the overall architecture of our proposed HANet.

A. *Progressive Foreground-Balanced Sampling*

We propose a PFBS approach that enables the model to accurately learn the features of the foreground image during the early training process. The basic idea behind PFBS is to improve the influence of changed pixels on model training, and moreover, to solve the sample imbalance problem by training the foreground image first and then the background image slowly thereafter. For ease of understanding, we here use the WHU-CD dataset for an example presentation, which contains a total of 4536 training set images (including 1200 foreground images and 3336 background images). As visualized in Fig. 3, there are four lines representing the four training processes of Normal Train, Fixed-X, Linear-Y, and Fixed-X Linear-Y. These are introduced in more detail as follows.

1) *Normal Train* means that all training datasets are used throughout the whole training process. Therefore, the model learns the entire dataset at each epoch, completely

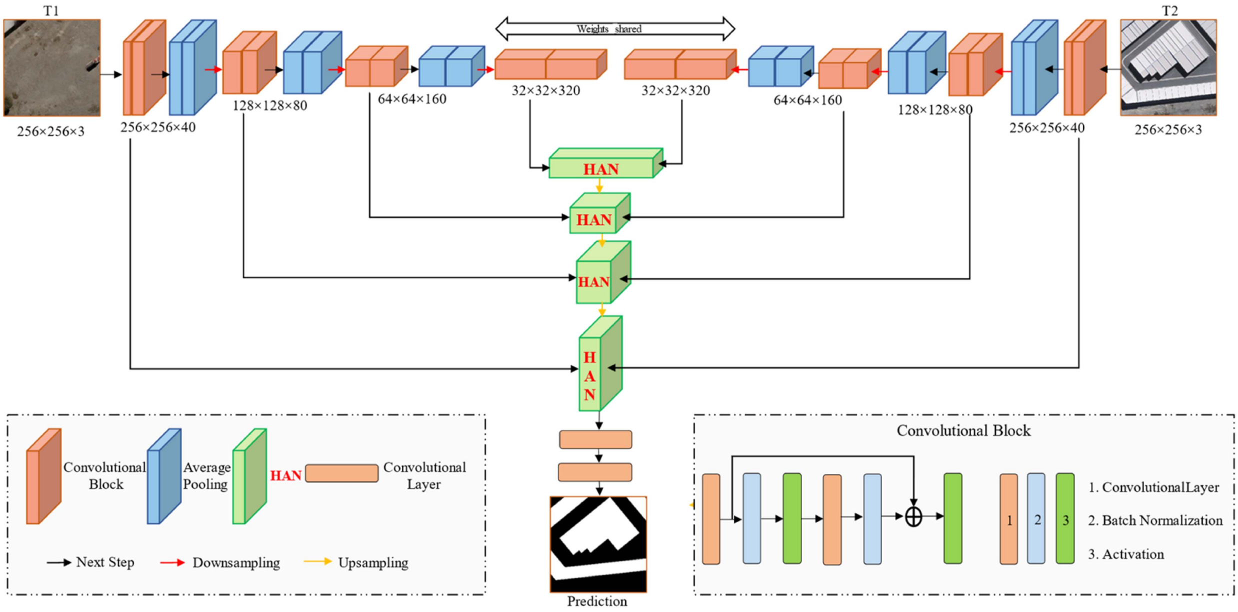

Fig. 4. Illustration of our proposed HANet model.

ignoring the fact that this binary dataset is extremely unbalanced.

2) *Fixed-X* means that the first X epochs are trained with 1200 foreground images, after which the whole training dataset is used for training. Notably, Fixed-X takes the data imbalance into consideration. It first helps the model to learn the “change” features by specifically learning the foreground images with changing information, and then enhances the robustness of the model by learning the entire dataset (including background images).

3) *Linear-Y* means that the first Y epochs are trained by gradually and linearly increasing the ratio of background images to foreground images, after which the whole training dataset is used for training. Similarly, Linear-Y also takes the imbalance of data into consideration. It only learns the features of the foreground images at the time of the first epoch, then gradually learns the features of the background images at the time of the second epoch to epoch Y, and then learns the features of the entire dataset after epoch Y+1. In this example, 222 background images are added per epoch.

4) *Fixed-X Linear-Y* refers to the combination of *Fixed-X* and *Linear-Y*. Under this approach, the model learns the foreground images at the first X epoch, then adds the background images linearly through Y epochs, and subsequently learns the entire dataset after X+Y epochs. In this example, 333 background images are added per epoch.

B. HANet Details

As shown in Fig. 4, HANet is a Siamese architecture that incorporates a weight-shared feature extractor. It consists of two parts: a multiscale feature extractor and the HAN module. The multiscale feature extractor can obtain multiscale building semantic information. For its part, the HAN module can gradually obtain semantic integration and refinement features of buildings. The deep-learning-based CD method generally takes two different temporal images (T1 and T2) as input images and produces one prediction image as the output image. The process is described in more detail in the following.

Our HANet model includes a convolutional block, adaptive average pooling layer, and HAN module, which refer the MSPSNet [15]. T1 and T2 are of the same size; the dimension of T1 is $W \times H \times C$, where W is the width of the input image, H is the height of the input image, and C is the number of input image channels. For the VHR images used as input pairs, we use the size $256 \times 256 \times 3$. In our HANet, we extract four-scale building features in four steps. Every scale feature can be described by $\{f_i^m | i = 1, 2, m = 1, 2, 3, 4\}$. Here, f_i^m means each temporal image, i represents T1 and T2, and m denotes the four different scales. When we get the first f_i^m, it will simultaneously be taken as the input of the HAN module. The HAN module can continuously help the model to identify the changed regional features and improve the feature details.

In general, our HANet architecture consists of four convolutional blocks and three adaptive average pooling layers. In a convolutional block, a residual connection is used to superimpose the feature. The order is as follows: the convolutional layer, batch normalization (BN) layer, and activation functions rectified linear unit (ReLU). The convolutional layer, BN, and ReLU repeat twice. Therefore, the result of the convolutional block can be denoted as follows:

$$g\left(f_i^m\right) = R\left(B_N\left(conv_1^{3\times3}\left(f_i^m\right)\right)\right) \tag{1}$$

$$Output_0 = R\Big(B_N\left(\left(conv_2^{1\times1}\left(g\left(f_i^m\right)\right)\right) + conv_1^{3\times3}\left(f_i^m\right)\right)\Big) \tag{2}$$

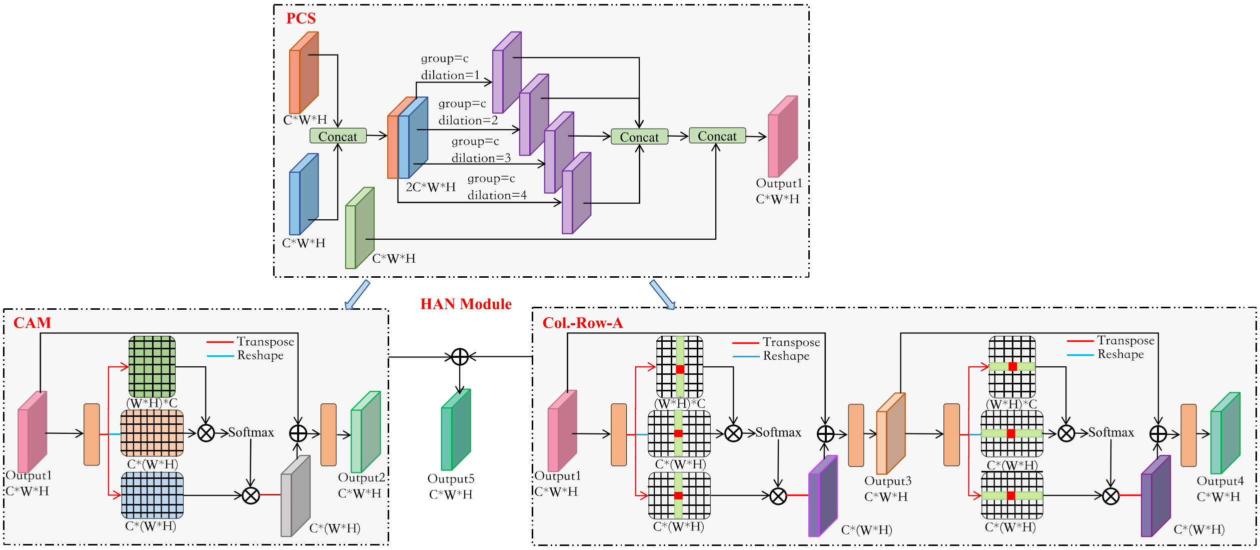

Fig. 5. Illustration of our HAN module. Here, HAN means hierarchical attention network, PCS means parallel convolutional structure, CAM means channel attention module, and Col.-Row-A means column attention and row attention.

where $\{f_i^m | i = 1, 2, m = 1, 2, 3, 4\}$ represents the input features. $conv_1^{3\times 3}$ represents the first convolutional layer with a kernel size of 3. $conv_2^{1\times 1}$ represents the second convolutional layer with a kernel size of 1. B_N is the normalization operation and R is the activation operation.

In the adaptive average pooling layer, we use 128-, 64-, and 32-layer features to obtain the multiscale feature blocks. Thus, our HANet can easily learn how to gradually and deeply extract semantic features, which makes it discriminative.

C. HAN Module

As shown in Fig. 5, our HAN module consists of a parallel convolution structure (PCS) [15] and spatial–spectral axial attention, and the latter contains both column and row attention (Col.-Row-A) [43], [44]. PCS can increase the effectiveness of feature extraction and reduce the model computational cost. The calculation amount of SA is second order, and Col.-Row-A can reduce the calculation amount and achieve higher calculation efficiency. The input first passes through PCS for multiscale feature fusion, then passes through the refinement of CAM and Col.-Row-A at the same time. Finally, the results of CAM and Col.-Row-A are added together to produce the results of the HAN module.

PCS has four different group convolutions [45] with a kernel size of 3 × 3 and four dilation parameters, enabling it to integrate multiscale features. Group convolution can expand the receptive fields. c is the number of group members in each group convolution operation, and is 0.5 times the size of the input data. Therefore, the PCS can be denoted as follows:

$$F = cat\left(f_1^m, f_2^m\right), m = 1, 2, 3, 4 \tag{3}$$

$$\begin{aligned} S = cat\left(conv_{d=1}^{3\times 3}(F), conv_{d=2}^{3\times 3}(F),\right. \\ \left. conv_{d=3}^{3\times 3}(F), conv_{d=4}^{3\times 3}(F)\right) \end{aligned} \tag{4}$$

$$Output_1 = conv^{1\times 1}(S) \tag{5}$$

where $cat(\cdot)$ represents the channel dimension concatenation operation, $conv_{d=m}^{3\times 3}$ is the convolutional layer with kernel size 3 and dilation of m, and $conv^{1\times 1}$ represents the 1 × 1 convolutional layer.

CAM [46], [47] can refine the detailed feature. It contains a 3 × 3 and a 1 × 1 convolutional layer, an elementwise sum operation, two matrix multiplication operations, a transpose operation, three reshape operations, and an activation function. Therefore, the CAM can be denoted as follows:

$$I_0 = conv_1^{3\times 3}(Output_1) \tag{6}$$

$$\begin{aligned} Output_2 = conv_2^{1\times 1}\left(S_o\left(R_e(I_0) \times R_e\left(T_r\left(I_0\right)\right)\right)\right. \\ \left. \times R_e\left(I_0\right) + I_0\right) \end{aligned} \tag{7}$$

where $conv_1^{3\times 3}$ is the initiatory convolutional layer with a kernel size of 3, whereas R_e represents the reshape operation. T_r represents the transpose operation, S_o is the SoftMax function, and $conv_2^{1\times 1}$ is the last convolutional layer with a kernel size of 1.

Col.-Row-A is very similar to CAM in implementation, with the major difference relating to the axial attention. The amount of computation for CAM is second-order, and Col.-Row-A can reduce the amount of computation required and achieve higher computational efficiency. Col.-Row-A is calculated first in the horizontal direction and then in the vertical direction to reduce the computational complexity. The receptive field of Col.-Row-A is W (or H) pixels in the same row (or column) of the target pixel, meaning that it has a much smaller receptive field than CAM. Therefore, the column attention can be denoted as follows:

$$I_1 = conv_1^{3\times 3}(Output_1) \tag{8}$$

$$\begin{aligned} Output_3 = conv_2^{1\times 1}\left(S_o\left(R_e^c(I_1) \times R_e^c\left(T_r\left(I_1\right)\right)\right)\right. \\ \left. \times R_e^c\left(I_1\right) + I_1\right) \end{aligned} \tag{9}$$

where R_e^c represents the reshape operation of column attention. The row attention can be denoted as follows:

$$I_2 = conv_1^{3\times 3}\left(Output_3\right) \tag{10}$$

$$\begin{aligned} Output_4 = conv_2^{1\times 1}\left(S_o\left(R_e^r(I_2) \times R_e^r\left(T_r\left(I_2\right)\right)\right)\right. \\ \left. \times R_e^r\left(I_2\right) + I_2\right). \end{aligned} \tag{11}$$

Here, R_e^r represents the reshape operation of row attention. Finally, we can create HAN module by adding CAM and Col.-Row-A

$$Output_5 = Output_2 + Output_4. \tag{12}$$

Through the continuous improvement with PCS, CAM and Col.-Row-A, context information, the interior, and the edge features of the building can be better extracted. In fact, our HAN module considers not only multiscale information extraction but also contextual information extraction, as well as the number of modules parameter.

D. Loss Function

For this extremely unbalanced CD challenge, we adopt hybrid loss [48] as the loss function to alleviate the impact of the data imbalance. Hybrid loss combines weighted cross-entropy loss and dice loss, and can be expressed as follows:

$$L = L_w + L_d \tag{13}$$

$$L_w = \frac{1}{W \times H}\sum_{i=1}^{W\times H} w\left[cla\right] \cdot \left(\log\left(\frac{exp\left(\hat{y}\left[i\right]\left[cla\right]\right)}{\sum_{l=0}^{1} exp\left(\hat{y}\left[i\right]\left[l\right]\right)}\right)\right) \tag{14}$$

$$L_d = 1 - \frac{2 \cdot S_o\left(\hat{Y}\right)}{Y + S_o\left(\hat{Y}\right)} \tag{15}$$

where L_w represents the weighted cross-entropy loss, L_d represents the dice loss, w represents the weights, the value of cla is either 1 or 0 (corresponding to changed and unchanged pixels, respectively), $\hat{y}[i]$ represents the ith point in $\hat{Y}$, i and l are indexes, $\hat{Y} = \{\hat{y}[i], i = 1, 2, \ldots, W \times H\}$ represents the change map, and $\hat{Y}$ represents the ground truth.

IV. Experiment

In this section, we introduce the experimental datasets and environment, comparison models, and evaluation metrics. We then discuss the experimental results and ablation study in detail.

A. Experimental Setup

WHU-CD [21] is a public RS building CD dataset, which contains one large VHR (0.2 m/pixel) image patch pair with a size of 32 507 ×15 345, which is cropped by 512 × 512. It includes areas of Christchurch in New Zealand, where a 6.3-magnitude earthquake occurred in 2011, after which significant rebuilding took place. There are 21 442 501 changed pixels, accounting for 4.26% of the total, and 481 873 979 unchanged pixels, accounting for 95.74%. It is therefore an extremely unbalanced binary classification dataset (the changed pixels in the label are 1, and the unchanged pixels are 0), as shown in Fig. 2. We adopted the default data split (training: 1260 image pairs; testing: 690 image pairs) published on the authors' websites. Due to GPU memory capacity limitations, and to facilitate a fair comparison with other algorithms, we directly cropped the default image patch pairs into sizes of 256 × 256 with no overlap. Also, we randomly selected 10% of images from the training dataset to form the validation dataset. Therefore, we obtained a dataset including 4536/504/2760 pairs of patches for training/validation/testing, respectively.

LEVIR-CD [13] is a public large-scale RS building CD dataset, which consists of 637 VHR (0.5 m/pixel) image patch pairs with a size of 1024 × 1024 pixels. It includes 20 different areas in several cities in the United States and contains a large number of seasonal and light changes, which makes CD more difficult. There are 31 066 643 changed pixels, accounting for 4.65% of the total, and 636 876 269 unchanged pixels, accounting for 95.35%. It is therefore an extremely unbalanced binary classification dataset (the changed pixels in the label are 1, and the unchanged pixels are 0), as shown in Fig. 2. We adopted the default data split (training: 445 image pairs; validation: 64 image pairs; testing: 128 image pairs) published on the authors' websites. Due to GPU memory capacity limitations, and to facilitate a fair comparison with other algorithms, we directly cropped the default image patch pairs into sizes of 256 × 256 with no overlap. Therefore, we obtained a dataset including 7120/1024/2048 pairs of patches for training/validation/testing, respectively.

Implementation details: Our models are implemented on PyTorch and trained using a single NVIDIA RTX 3090 GPU. We adopt the Adam optimizer with a weight decay 5e–4 and learning rate 5e–4 (with the gamma of 0.5 adopted to update the learning rate) to minimize the loss. We use StepLR with a step size of 8 and gamma of 0.5 to update our learning rate, as shown in (16). Due to the limitations of the GPU, we set a batch size of 8 and epoch number of 100 to make the model converge. We trained our model for 100 epochs and saved the best model on the validation set as the final training result. We refer to the fixed foreground image training method as *Fixed-X* and the linear foreground image increase method as *Linear-X*; here, X means the number of epochs

$$New_{lr} = initial_{lr} \times \gamma^{epoch//step\ size}. \tag{16}$$

Evaluation metrics: In order to quantitatively verify the effectiveness of our proposed model, we use the following metrics for evaluation: F1-score ($F1$), Precision ($Pre.$), Recall ($Rec.$), Overall Accuracy (OA), and Kappa Coefficient (KC). These are employed by comparing the GT and prediction maps, and can be specifically defined as follows:

$$F1 = \frac{2}{Pre.^{-1} + Rec.^{-1}} \tag{17}$$

$$Pre. = TP/\left(TP + FP\right) \tag{18}$$

$$Rec. = TP/\left(TP + FN\right) \tag{19}$$

$$OA = \left(TP + TN\right)/\left(TP + TN + FN + FP\right) \tag{20}$$

$$KC = \frac{OA - PRE}{1 - PRE} \tag{21}$$

$$IoU = TP/(TP + FN + FP) \tag{22}$$

$$PRE = \frac{(TP + FN) \times (TP + FP)}{(TP + TN + FP + FN)^2} + \frac{(TN + FP) \times (TN + FN)}{(TP + TN + FP + FN)^2} \tag{23}$$

where TP denotes the number of true positives, TN denotes the number of true negatives, FP denotes the number of false positives, and FN denotes the number of false negatives. “PRE” represents the sum of the “ground truth and the product of the predicted result” corresponding to all categories, divided by the “average of the total sample data.” It is worth noting that higher values of F1, OA, and KC indicate good CD performance.

B. Comparison With State-of-the-Art Methods

It is necessary to compare our proposed model with three different types of state-of-the-art methods (pure CNN-based methods, attention-based methods, and transformer-based methods). Accordingly, we choose three classic pure CNN-based models (FC-EF [11], FC-Siam-conc [11], and FC-Siam-diff [11]), three attention-based models (STANet [13], SNUNet [14], and MSPSNet [15]), and three transformer-based models (BIT [17], Change Former [18], and RSP-BIT [19]).

1) *FC-EF [11]:* A fully convolutional early fusion network, which is directly based on the U-Net structure, that concatenates two input images before feeding them into the network, then treats them as different channels of an image.
2) *FC-Siam-conc [11]:* A fully convolutional Siamese-concatenation model, which skip-connects the three feature diagrams from the two encoder branches and the corresponding layer of the decoder.
3) *FC-Siam-diff [11]:* A fully convolutional Siamese-difference model, which first obtains the absolute value of the difference between the two decoder branches, after which a skip connection is made with the corresponding layer of the decoder.
4) *STANet [13]:* A Siamese-based spatial–temporal attention neural network, which can get more discriminative features through the use of the spatial–temporal attention module.
5) *SNUNet [14]:* A densely connected Siamese network, which alleviates the loss of localization information in the deep layers of neural networks and refines the most representative features of different semantic levels via ensemble channel attention module.
6) *MSPSNet [15]:* A deep multiscale Siamese network, which obtains the multiscale feature via PCS and SA.
7) *BIT [17]:* A bitemporal image transformer network, which can efficiently and effectively model contexts within the spatial–temporal domain by using the transformer to capture the contextual information between different temporal images.
8) *Change Former [18]:* A transformer-based Siamese network architecture, which unifies a hierarchically structured transformer encoder with a multilayer perception decoder in its Siamese network architecture to efficiently render multiscale long-range details.
9) *RSP-BIT [19]:* A BIT model with remote sensing pretraining (RSP), which adopts the ViTAE [49] network and uses the MillionAID [50] dataset for pretraining.

Notably, we implement these CD models using their public codes with default hyperparameters found on GitHub in the same software environment.

Table I shows the specific quantitative comparison results on the WHU-CD and LEVIR-CD datasets obtained by these methods. It can be observed that our HAN module is a lightweight and effective SA mechanism by calculating the parameters and floating-point operations (FLOPs) of the model. By observing the data in red, we can see that our proposed HANet achieves good performance. Moreover, on the key metrics of F1-score, OA, and KC, HANet also achieves better performance. In particular, the WHU-CD and LEVIR-CD datasets are among the more unbalanced samples from Fig. 2. For example, the F1-score of our HANet exceeds that of Change Former by 0.98/0.08 points on these two datasets, respectively. Getting such a result is a little difficult because Change Former is a transformer-based method, which has more network parameters. Note that our HANet is better at integrating contextual semantic information at different scales. This may be attributed to the powerful semantic information extraction ability of our HAN module.

Fig. 6 presents the visualization comparison results on the WHU-CD and LEVIR-CD datasets with these methods. We selected some challenging samples (a)–(h) with complex building features from the test set for illustrative purposes. This diagram can be used to intuitively compare the performance of each model. To further improve the intuitiveness of this visualization, we use several colors and a small red box to show the key and detailed results of the model. Blue and red colors indicate missed detection and error detection, respectively.

First, our HANet is better able to avoid false negatives [e.g., Fig. 6(a), (b), (g), and (h)] on complex building features. For example, the pure CNN-based FC-EF and FC-Siam-conc methods have obviously missed detections. Compared with the attention-based SNUNet and MSPSNet on (d), it can be easily seen that our proposed HANet has fewer blue pixels. Moreover, when compared with the transformer-based methods BIT, Change Former, and RSP-BIT in Fig. 6(c), HANet also demonstrates good performance when extracting small building semantic features.

Second, our HANet is better at avoiding false positives [e.g., Fig. 6(b), (d), and (g)] on complex background features, including the irrelevant changes caused by seasonal variation and land use cover change. By observing the red area next to the building, we can determine that FC-EF, STANet, SNUNet, and MSPSNet are more sensitive to the changes in certain trees, which are not buildings. In Fig. 6(d), the U-shaped building undergoes a change in the impermeable material surface; our HANet is much better at extracting semantic changes in the region of interest.

Third, our HANet is better at extracting detailed building features [e.g., Fig. 6(h)]. In some nonconventional quadrilateral

TABLE I

RESULTS OF COMPARISON WITH OTHER SOTA CD METHODS IN TERMS OF PARAMETERS (PARA.), FLOPS, F1-SCORE, PRECISION, RECALL, OVERALL ACCURACY, AND INTERSECTION OVER UNION ON THE WHU-CD AND LEVIR-CD DATASETS

Model	Para.	FLO	WHU-CD					LEVIR-CD				
	(M)	Ps(G)	F1	Pre.	Rec.	OA	IoU	F1	Pre.	Rec.	OA	IoU
FC-EF [11]	1.35	-	58.05	76.49	46.77	-	40.89	61.52	73.31	53.00	-	44.43
FC-Siam-conc [11]	1.54	2.29	63.99	72.06	57.55	-	47.05	64.41	95.30	48.65	-	47.51
FC-Siam-diff [11]	1.35	2.29	86.31	89.63	83.22	-	75.91	89.00	91.76	86.40	-	80.18
STANet-PAM [13]	16.93	6.58	82.00	75.70	**89.30**	98.60	69.44	85.20	80.80	90.10	98.40	74.22
SNUNet [14]	12.03	27.44	87.76	87.84	87.68	**99.13**	78.19	**89.97**	91.31	88.67	**98.99**	**81.77**
MSPSNet [15]	2.21	14.17	86.49	87.84	85.17	99.05	76.19	89.67	90.75	88.61	98.96	81.27
BIT [17]	3.55	4.35	80.97	74.01	89.37	98.51	68.02	89.94	90.33	89.56	98.98	81.72
Change Former [18]	20.75	-	**87.18**	92.70	82.28	99.14	77.27	90.20	92.05	88.37	99.01	**82.21**
RSP-BIT [19]	24.44	-	78.50	69.93	89.45	98.26	64.60	89.71	**92.00**	87.53	98.98	81.34
HANet(ours)	3.03	14.07	88.16	**88.30**	88.01	99.16	78.82	90.28	91.21	**89.36**	99.02	82.27

All values are in %. Higher values of F1 and OA indicate good CD performance. For convenience: Best, 2nd-best, and 3rd-best.

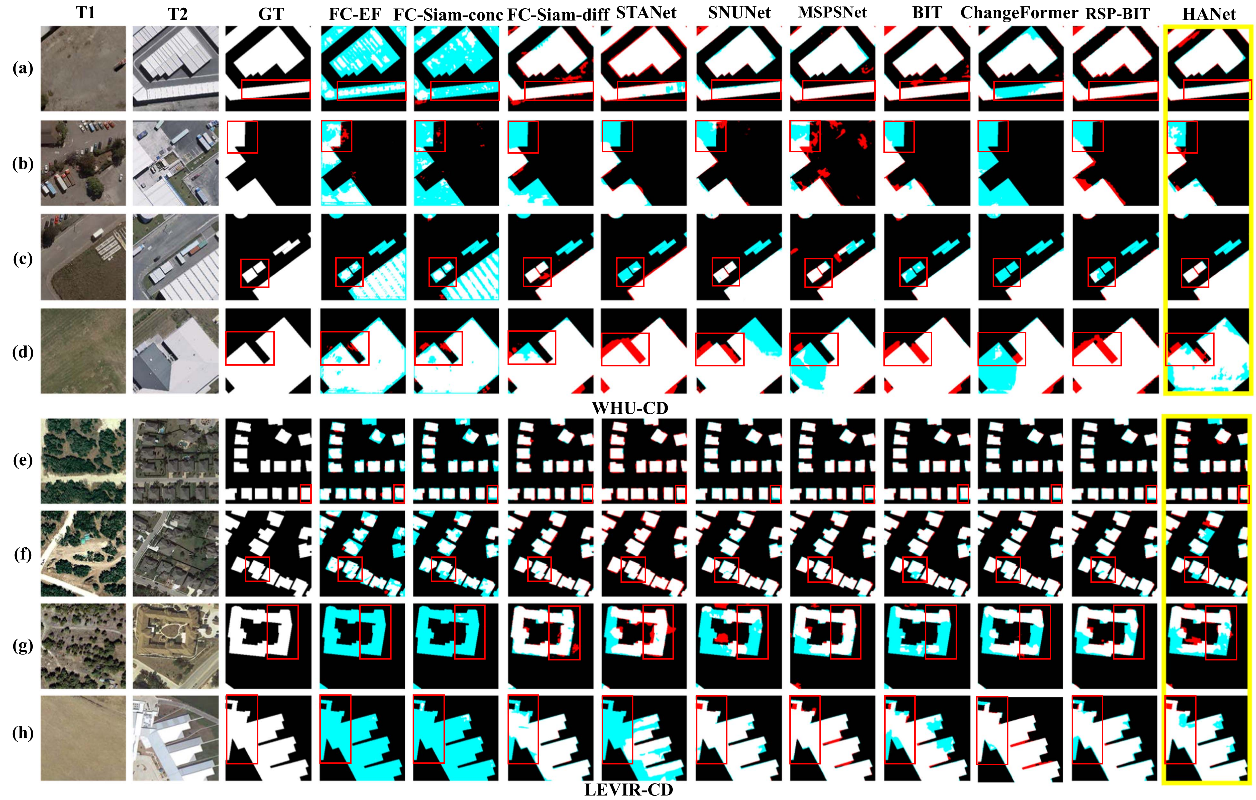

Fig. 6. Visualization comparison of different methods on the WHU-CD and LEVIR-CD test sets. For convenience, several colors are used to facilitate a clearer visualization of results, i.e., TP (white), FP (red), TN (black), and FN (blue).

buildings, even if the building size is very large, our HANet is still better at extracting the edge features of irregular buildings.

C. Ablation Study

We set up the following models to validate the effectiveness of our proposed model.

1) *Base:* The baseline model consists of the CNN Siamese network with PCS and CAM.
2) *HANet:* Base model + Row-Col.-A.
3) *HANet-Fixed-X:* HANet + Fixed-X.
4) *HANet-Linear-X:* HANet + Linear-X.
5) *HANet-Fixed-X Linear-Y:* HANet + Fixed X and then Linear-Y.

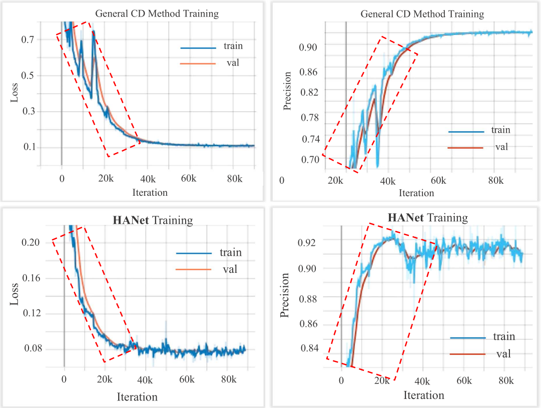

Fig. 7. Visualization of general CD method's and HANet's training process with regard to loss and precision.

TABLE II
ABLATION STUDY ON THE FIXED-X (F-X) OF PFBS ON WHU-CD

Model	Fixed-X on WHU-CD						
	Base	+ F-5	+ F-15	+ F-25	F1	Pre.	Rec.
Base	√				84.36	90.44	81.03
+F-5		√			86.68	85.45	87.96
+F-15			√		**88.49**	**88.99**	**88.00**
+F-25				√	88.27	90.25	86.37

F1-Score, Precision, and Recall are compared.

There are several hyperparameters in PFBS and attention modules in the HAN module. Therefore, we choose some of them for an ablation study. We select some of the typical ablation study results for presentation in order to intuitively observe the performance. Here, the white area indicates prediction. Special attention should be paid to the four-color box: here, each color stands for a different kind of typical ablation study.

Ablation on training stability: Visualizing the training process of general CD method and HANet helps us to improve our understanding of the training process. From Fig. 7, we can observe that the general CD method has an unstable process during convergence, around 20 K iterations (equal to 20 epochs). After adding PFBS, HANet convergence becomes smooth, which means PFBS has a good effect. Under these circumstances, HANet finds it easier to converge.

Ablation on PFBS: In the Fixed-X of PFBS, we discussed the hyperparameter of epoch numbers. As shown in Table II, we conducted an experiment at intervals of five epochs from 5 to 40 epochs. We present the higher results of the 5, 15, and 25 epochs. Regardless of which parameter we use, the results exceed the baseline. Moreover, the training result of Fixed-15 reached 88.49% in terms of the F1-score, which is a parameter with the best performance.

As a result, as shown in Fig. 8, Fixed-15 also achieves better performance than the baseline. This shows that the building semantic features in the foreground images of WHU-CD can be learned in 15 epochs. This result shows that different datasets should have different standards in this PFBS.

TABLE III
ABLATION STUDY ON THE FIXED-X (F-X) AND LINEAR-X (L-X) OF PFBS ON WHU-CD

Model	Fixed-X + Linear-X on WHU-CD					
	Ba se	+ L-25	+F-10 L-10	+F-15 L-15	F1	Pre.
Base	√				84.36	90.44
+L-25		√			87.18	89.79
+F-10 L-10			√		**88.10**	**88.78**
+F-15 L-15				√	86.41	86.82

F1-Score, Precision, and Recall are compared.

TABLE IV
ABLATION STUDY ON THE KEY PART OF HAN MODULE (COL.-ROW-A, ABBREVIATED TO C./R) ON WHU-CD

Model	Row/Col.-Attention on WHU-CD						
	Base	+ Col.	+ Row	+ C./R	F1	Pre.	Rec.
Base	√				84.36	90.44	81.03
+Col.		√			87.31	87.84	86.78
+Row			√		87.73	88.29	87.18
+C./R				√	**87.78**	**87.45**	**88.12**

F1-Score, Precision, and Recall are compared.

Similarly, we discuss the hyperparameters of Linear-X and the combination of Fixed-X and Linear-Y. As shown in Table III, regardless of which parameters are employed, the results of our approach can exceed the baseline. Note that there are many possible combinations of X and Y. Experiments have shown that the use of more foreground images for training does not necessarily produce better performance; in fact, the performance depends on the scale of the foreground image. As shown in Fig. 8, Fixed-10 plus Linear-10 yields relatively good results.

Ablation on HAN module: In the HAN module, the key component is attention, which enables contextual semantic information about the building to be obtained. As shown in Table IV, we investigate the performance of axial attention, including column attention and row attention. It is obvious that single-column attention or row attention produces different scores because the model initializes parameters randomly. Experiments show that the combination of column and row attention is superior to single axial attention.

Ablation on PFBS and HAN module: We next discuss the combination of our proposed PFBS and HAN module. As shown in Table V, the performance of both Linear-X and Fixed-X plus HAN module exceed the baseline. Moreover, the performance of Fixed-X plus HAN module is better than that of Linear-X

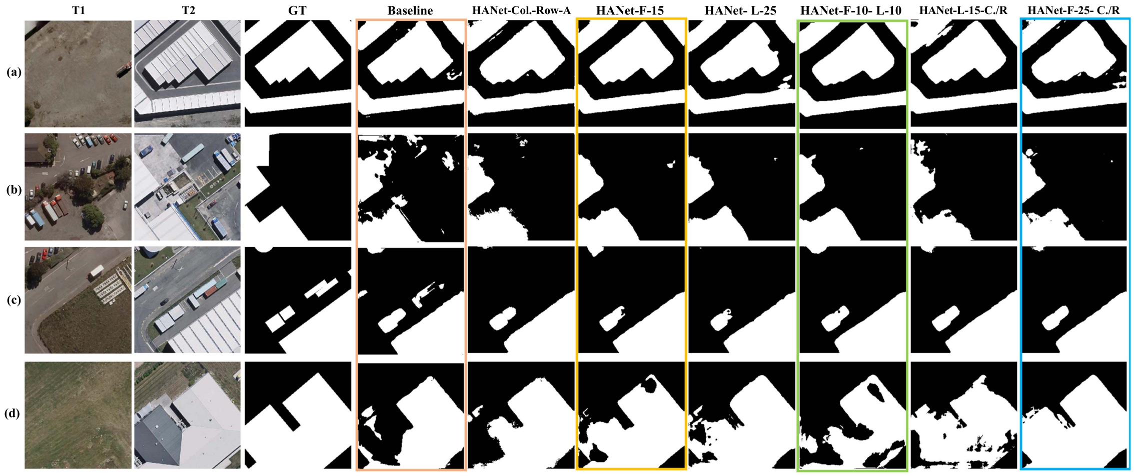

Fig. 8. Visualization of ablation study results.

TABLE V
Ablation Study on the Fixed-X (F-X) and Linear-X (L-X) of PFBS and the Key Part of the HAN Module (Col.-Row-A, Abbreviated to C./R) on WHU-CD

Model	Row/Col.-A-Fixed-X + Linear-X on WHU-CD						
	Base	+ L-15 C./R	+ L-25 C./R	+ F-15 C./R	F1	Pre.	Rec.
Base	√				84.36	90.44	81.03
+L-15 C./R		√			87.57	87.35	87.78
+L-25 C./R			√		87.26	89.15	85.45
+F-25 C./R				√	**88.16**	**88.30**	**88.01**

F1-Score, Precision, and Recall are compared.

TABLE VI
Ablation Study of PFBS on MSPSNet

Model	WHU-CD					
	F1	Pre.	Rec.	OA	KC	IoU
MSPSNet [15]	86.49	87.84	85.17	99.05	86.00	76.19
+F-10	87.50	87.66	87.34	99.11	87.04	77.77
+L-15	86.88	87.63	86.14	99.08	86.40	76.80
+F-5 L-10	**87.65**	**87.25**	**88.05**	**99.12**	**87.19**	**78.01**

TABLE VII
Ablation Study of the Model With Different Loss Functions Including Hybrid Loss (HL) and Focal Loss (FL) on WHU-CD

Loss Model	Different loss functions on WHU-CD				
	HL	FL	HL+ **F-15**	FL+ **F-15**	F1
Hybrid Loss	√				84.36
Focal Loss		√			85.50
Hybrid Loss+ F-15			√		**88.49**
Focal Loss+ F-15				√	87.38

plus HAN module. From Fig. 7, we can observe that Fixed 25 plus Col.-Row-A is better than Linear-15 plus Col.-Row-A.

Ablation of PFBS on another method: We add our proposed PFBS into MSPSNet [15] to validate the efficiency of PFBS. As shown in Table VI, the three methods of PFBS (including Fixed-X, Linear-X, and the combination of Fixed-X and Linear-X) all work well on MSPSNet. Moreover, the performance of the combination of Fixed-X and Linear-X is best. Therefore, we conclude that it is possible to use PFBS to improve the performance of another method.

Ablation of different loss functions: In order to verify the effectiveness of PFBS, we compare the performance of different loss functions on WHU-CD. In this article, we use Focal Loss [51] and Hybrid Loss to compare. As shown in Table VII, we can know that the performance of Focal Loss is better than Hybrid Loss. Subsequently, our proposed PFBS (here we use Fixed-15) simultaneously improves the F1-score, whereas Hybrid Loss with Fixed-15 is better.

Ablation of different sampling strategies: There are some existing methods to address the class imbalance in CD except for using the loss function. For example, synthesize CD samples [23] that are more class-balanced or transfer a pretrained model [24] that is more robust. RSP-BIT [19] is a CD method for RSP on the MillionAID dataset. Table VIII shows that our performance is best when we use 100% of the training dataset, and SaDL and CDNet +IAug are better when we use 5% and 20% of the training dataset. We further analyze the reason and find that their method increases the number of pixels with change information, whereas our PFBS method improves the

TABLE VIII
ABLATION STUDY OF SAMPLING STRATEGIES ON LEVIR-CD

Sampling strategies	LEVIR-CD		
	100%	20%	5%
	F1		
RSP-BIT [19]	89.71	74.73	55.62
SaDL [24]	88.74	87.25	**79.44**
CDNet+IAug [23]	89.00	**87.50**	76.00
HANet-PFBS(F-15)	**90.28**	81.36	68.66

performance of the model without increasing the number of pixels with change information.

V. Conclusion

In this article, a PFBS approach on the basis of not adding change information is proposed to help the model accurately learn the features of the foreground image during the early stages of the training process. A discriminative Siamese network, HANet, is proposed to integrate multiscale features and refine detailed features. Extensive experimental validation on two extremely unbalanced binary CD datasets (WHU-CD and LEVIR-CD) shows that our proposed methods (PFBS and HANet) outperform many remarkable existing models. In future work, it would be worthwhile to address some problems with the existing PFBS, for example, the optimal solution of X and Y in Fixed-X and Linear-Y, more forms of PFBS (such as the method of nonlinear increase), and the performance of PFBS on additional models. Similarly, the HAN module in HANet is suitable for migration to test performance on other models.

References

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989, doi: 10.1080/01431168908903939.

[2] P. K. Mishra, A. Rai, and S. C. Rai, "Land use and land cover change detection using geospatial techniques in the Sikkim Himalaya, India," *Egypt. J. Remote Sens. Space Sci.*, vol. 23, no. 2, pp. 133–143, Aug. 2020, doi: 10.1016/j.ejrs.2019.02.001.

[3] G. Xian and C. Homer, "Updating the 2001 National Land Cover Database impervious surface products to 2006 using Landsat imagery change detection methods," *Remote Sens. Environ.*, vol. 114, no. 8, pp. 1676–1686, Aug. 2010, doi: 10.1016/j.rse.2010.02.018.

[4] C. Song, B. Huang, L. Ke, and K. S. Richards, "Remote sensing of alpine lake water environment changes on the Tibetan Plateau and surroundings: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 92, pp. 26–37, Jun. 2014, doi: 10.1016/j.isprsjprs.2014.03.001.

[5] P. Lu, Y. Qin, Z. Li, A. C. Mondini, and N. Casagli, "Landslide mapping from multi-sensor data through improved change detection-based Markov randomfield," *Remote Sens. Environ.*, vol. 231, Sep. 2019, Art. no. 111235, doi: 10.1016/j.rse.2019.111235.

[6] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[7] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, Apr. 1998, doi: 10.1016/S0034-4257(97)00162-4.

[8] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[9] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

[10] L. Wan, Y. Xiang, and H. You, "A post-classification comparison method for SAR and optical images change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1026–1030, Jul. 2019.

[11] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," Oct. 2018. Accessed: Oct. 04, 2022. [Online]. Available: http://arxiv.org/abs/1810.08462

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," May 2016. Accessed: Dec. 10, 2022. [Online]. Available: http://arxiv.org/abs/1409.0473

[13] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, May 2020, Art. no. 1662, doi: 10.3390/rs12101662.

[14] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 8007805, 10.1109/LGRS.2021.3056416.

[15] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, "Deep multiscale Siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2021, Art. no. 5406512, doi: 10.1109/TGRS.2021.3131993.

[16] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, p. 11.

[17] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2021, Art. no. 5607514, doi: 10.1109/TGRS.2021.3095166.

[18] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," Sep. 2022. Accessed: Dec. 10, 2022. [Online]. Available: http://arxiv.org/abs/2201.01293

[19] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," May 2022. Accessed: Dec. 10, 2022. [Online]. Available: http://arxiv.org/abs/2204.02825

[20] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-2, pp. 565–571, May 2018, doi: 10.5194/isprs-archives-XLII-2-565-2018.

[21] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[22] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2021, Art. no. 5604816, doi: 10.1109/TGRS.2021.3085870.

[23] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2021, Art. no. 5603216, doi: 10.1109/TGRS.2021.3066802.

[24] H. Chen, W. Li, S. Chen, and Z. Shi, "Semantic-aware dense representation learning for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5630018, doi 10.1109/TGRS.2022.3203769.

[25] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for detecting oriented objects in aerial images," Dec. 2018. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/1812.00155

[26] Y. Xu, B. Du, L. Zhang, and S. Chang, "A low-rank and sparse matrix decomposition-based dictionary reconstruction and anomaly extraction framework for hyperspectral anomaly detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1248–1252, Jul. 2020.

[27] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604–1617, Feb. 2021.

[28] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, "Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112589, doi: 10.1016/j.rse.2021.112589.

[29] Y. Xu and P. Ghamisi, "Consistency-regularized region-growing network for semantic segmentation of urban scenes with point-level annotations," *IEEE Trans. Image Process.*, vol. 31, pp. 5038–5051, 2022, doi: 10.1109/TIP.2022.3189825.

[30] M. Hu, C. Wu, and L. Zhang, "HyperNet: Self-supervised hyperspectral spatial–spectral feature understanding network for hyperspectral change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5543017, doi: 10.1109/TGRS.2022.3218795.

[31] H. Chen, C. Wu, B. Du, and L. Zhang, "Change detection in multi-temporal VHR images based on deep Siamese multi-scale convolutional networks," Jul. 2020. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/1906.11479

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," Mar. 2015. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/1411.4038

[33] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

[34] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 214–217.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," May 2015. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/1505.04597

[36] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, Jun. 2019, Art. no. 1382, doi: 10.3390/rs11111382.

[37] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," Jul. 2018. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/1807.10165

[38] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," Apr. 2016. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/1502.03044

[39] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, vol. 1, pp. 577–585.

[40] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020, doi: 10.1016/j.isprsjprs.2020.06.003.

[41] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2021, Art. no. 4403718, doi: 10.1109/TGRS.2021.3091758.

[42] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

[43] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," Dec. 2019. Accessed: Dec. 10, 2022. [Online]. Available: http://arxiv.org/abs/1912.12180

[44] J. Fu et al., "Dual attention network for scene segmentation," Apr. 2019. Accessed: Dec. 10, 2022. [Online]. Available: http://arxiv.org/abs/1809.02983

[45] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving CNN efficiency with hierarchical filter groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5977–5986.

[46] D. He, Q. Shi, X. Liu, Y. Zhong, G. Xia, and L. Zhang, "Generating annual high resolution land cover products for 28 metropolises in China based on a deep super-resolution mapping network using Landsat imagery," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 2036–2067, Dec. 2022, doi: 10.1080/15481603.2022.2142727.

[47] D. He, Q. Shi, X. Liu, Y. Zhong, and L. Zhang, "Generating 2m fine-scale urban tree cover product over 34 metropolises in China based on deep context-aware sub-pixel mapping network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, Feb. 2022, Art. no. 102667, doi: 10.1016/j.jag.2021.102667.

[48] L. Zhang, S. Zhou, J. Guan, and J. Zhang, "Accurate few-shot object detection with support-query mutual guidance and hybrid loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 14419–14427.

[49] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "ViTAE: Vision transformer advanced by exploring intrinsic inductive bias," Dec. 2021. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/2106.03348

[50] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-AID," Mar. 2021. Accessed: Dec. 11, 2022. [Online]. Available: http://arxiv.org/abs/2006.12485

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

Chengxi Han (Student Member, IEEE) received the B.S. degree in remote sensing science and technology from the School of Geosciences and Info-Physics, Central South University, Changsha, China, in 2018. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

He was a Trainee with the United Nations Satellite Centre of the United Nations Institute for Training and Research. His research interests include deep learning and remote sensing image change detection.

Mr. Han has been the IEEE GRSS Wuhan Student Branch Chapter Chair since 2021. He was the recipient of the IEEE GRSS 2022 Student Chapter Excellence Award.

Chen Wu (Member, IEEE) received the B.S. degree in surveying and mapping engineering from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote sensing, Wuhan University, Wuhan, China, in 2015.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. His research interests include multitemporal remote sensing image change detection and analysis in multispectral and hyperspectral images.

Haonan Guo (Student Member, IEEE) received the B.S. degree in geographical information science from Sun Yat-sen University, Guangzhou, China, in 2020. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include deep learning, building footprint extraction, urban remote sensing, and multisensor image processing.

Meiqi Hu (Graduate Student Member, IEEE) received the B.S. degree in surveying and mapping engineering from the School of Geoscience and Info-Physics, Central South University, Changsha, China, in 2019. She is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

Her research interests include deep learning and multitemporal remote sensing image change detection.

Hongruixuan Chen (Student Member, IEEE) received the B.E. degree in surveying and mapping engineering from the School of Resources and Environmental Engineering, Anhui University, Hefei, China, in 2019, and the M.E. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2022. He is currently working toward the Ph.D. degree in complexity science and engineering with the Graduate School of Frontier Science, The University of Tokyo, Chiba, Japan.

He was a Trainee with the United Nations Satellite Centre of the United Nations Institute for Training and Research. His research interests include deep learning, domain adaptation, and multimodal remote sensing image interpretation and analysis.

Mr. Chen is a Reviewer for eight international journals, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.