LRDE-Net: Large Receptive Field and Image Difference Enhancement Network for Remote Sensing Images Change Detection

Lele Li, Liejun Wang, Anyu Du, and Yongming Li

(Special Section: Remote Sensing for Smart Cities: Datasets, Algorithms and Applications)

***Abstract*—In the field of remote sensing, change detection is a crucial study area. Deep learning has made significant strides in the study of remote sensing image change detection during the past few years. Deep learning techniques still have some drawbacks. The global context cannot be modeled by convolutional neural networks due to the receptive field's restrictions. When extracting visual characteristics, the neural network does not concentrate more on the change region, which results in poor distinction between change and no-change regions. To address these problems, we propose networks with large receptive fields (LRFs) and difference image enhancement. First, we design the LRF strategy. It employs a long kernel shape in one spatial dimension for obtaining a long range of relations. Keeping a narrow kernel size in the other spatial dimension can extract local context information while avoiding interference from irrelevant regions. To focus on the changing features, we design the image difference enhancement (IDE) method, which decreases the distance between invariant features and enlarges the distance between changing features. In addition, we design the cross-channel interaction (CNI) strategy, which models the relationship between feature map channels and extracts feature representations through local CNI. On the CDD, WHU-CD, and LEVIR-CD public datasets, we conducted comprehensive experiments. According to the experimental results, our proposed LRDE-Net performs better than other state-of-the-art change detection techniques, and the change regions are more precisely identified. It can better cope with seasonal changes, light intensity, and other pseudochange disturbances.**

***Index Terms*—Change detection, convolutional neural network, cross-channel interaction (CNI), image difference enhancement (IDE), large receptive field (LRF), remote sensing.**

I. Introduction

IN ORDER to discover changes in an area, change detection involves recognizing and analyzing remote sensing photos taken at several periods in the same region. Change detection has important applications in urban planning, agricultural surveys, disaster warning and assessment, and ecosystem monitoring. With the advancement of optical sensors, many high-resolution remote sensing pictures may be collected and used for the detection of surface change. A key area of study in the field of remote sensing is the accurate and effective extraction of useful change information from a large collection of high-resolution photographs.

The existing approaches for change detection have mostly consisted of two categories: 1) algebraic operation-based methods [1], [2]; 2) image transformation-based methods [3], [4]. These techniques primarily make use of manually created characteristics to extract information about changes from the bitemporal images. Nevertheless, the features that are developed manually exhibit limited robustness and lack sophisticated semantic information, restricting their applicability solely to change detection in low-resolution remote sensing images [5]. The development of satellite sensing technology has led to the use of high-resolution remote sensing pictures for a range of remote sensing image change detection approaches. High-resolution remote sensing photos provide a wealth of characteristics and advanced semantic data. Due to the limited accuracy of conventional change detection approaches, it is difficult to make a significant advancement in high-resolution remote sensing picture change detection.

Global semantic information and contextual linkages of features at different layers cannot be fully retrieved due to the convolutional neural network's limited receptive field. In [6], dilated convolution is utilized to increase the convolutional neural network's receptive field. Contextual information may be systematically aggregated via dilated convolution. The pyramid pooling module is employed in [7] to collect features with various receptive fields. It obtains information about the variations between different scales and different subregions. Dilated convolution and pyramid pooling have a common drawback that they can only obtain information within a square window, whereas there is a variety of dissimilar contextual information in the real scene. The change areas in a remote sensing image change detection scenario may be long bands or discretely scattered, as shown in Fig. 1. We have chosen some pictures as examples for illustration. Long strip features are shown as roads in the red boxes in Fig. 1(a) and (b). Discrete features are shown as buildings in the green boxes in Fig. 1(c) and (d).

Manuscript received 30 August 2023; revised 3 October 2023; accepted 17 October 2023. Date of publication 23 October 2023; date of current version 23 November 2023. This work was supported in part by the Scientific and technological innovation 2030 major project under Grant 2022ZD0115800, and in part by theXinjiang Uygur Autonomous Region Tianshan Excellence Project under Grant 2022TSYCLJ0036. *(Corresponding author: Liejun Wang.)*

The authors are with the School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China (e-mail: lewis268@stu.xju.edu.cn; wljxju@xju.edu.cn; anydxju@xju.edu.cn; lym@xju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3326962

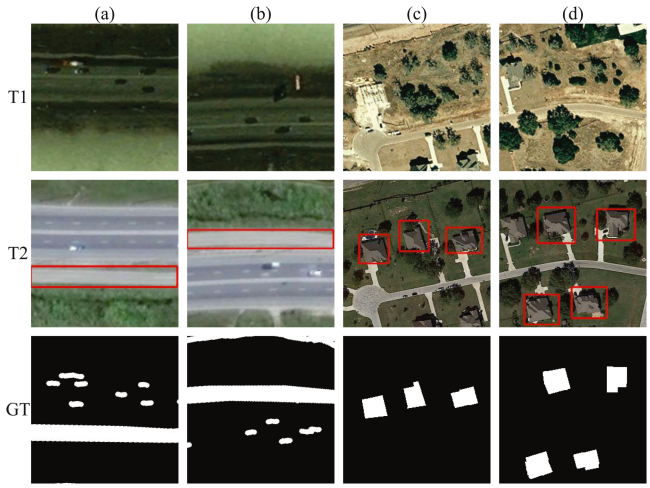

Fig. 1. Example of long strip or discrete distribution of change regions in remote sensing. T1 and T2 represent images acquired for the same location at the T1 and T2 moments, respectively. GT represents the labeling of the image that has changed at the T2 moment compared to the T1 moment. (a)–(d) denote 4 different columns of images, each consisting of 2 images taken at different moments and change map labels, where (a) and (b) denote images containing long strip of change regions, and (c) and (d) denote images containing discrete distribution of change regions.

These features correspond to the change labels in the third row of Fig. 1. The use of dilated convolution or pyramid pooling will bring in disturbing information from irrelevant regions. To solve this problem, we propose the large receptive field (LRF) strategy. In one spatial dimension, it uses a lengthy kernel shape to achieve a wide variety of relations in isolated regions. Second, by keeping a narrow kernel size in another spatial dimension, it can capture local context information, avoiding interference from irrelevant regions.

As remote sensing sensor technology advances, we can provide clear photos with better resolution and more detailed content. It also contains pseudochange interference information, such as light change and seasonal change. When extracting image features, the Siamese network method inputs two moments of images at the same time and does not focus on the change region. The result is that the distinction between changing and unchanging regions is not too strong, which can cause some irrelevant changes to interfere with the detection results [8]. In order for the network to obtain the intrinsic characteristics of truly changing regions, more research is needed to extract more discriminative difference information. In order to tackle this issue, we designed the image difference enhancement (IDE) strategy. This approach aims to reduce the distances among invariant features and amplify the gap among changing features. As a consequence, it obtains more discriminative detection results.

The following are the primary contributions of this article:

1) A novel remote sensing image change detection model, LRDE-Net, has been proposed, which includes the LRF strategy, the cross-channel interaction (CNI) strategy, and the IDE strategy. It can obtain multiscale feature representations and rich contextual information, obtain more discriminative change detection results, and effectively eliminate change interference information, such as lighting changes and seasonal changes.
2) The LRF strategy is designed to address the issue that the convolutional neural network experiences the receptive field's constraints and is unable to acquire global semantic information. It employs a long kernel shape in one spatial dimension for obtaining global contextual information and expanding the convolutional neural network's receptive field. In addition, we designed the CNI strategy to simulate the interaction between channels, so as to obtain an improved feature representation.
3) The IDE strategy was designed to address the issue that the network model has adequate discriminative capabilities and is prone to giving some erroneous detection results. By giving the change feature greater weight, it increases the distance between change pixels and obtains more discriminative change detection results.
4) On three open datasets, we ran a significant number of quantitative and qualitative studies. The results of the experiments show that our proposed strategy outperforms eight SOTA approaches by a wide margin.

The rest of this article is organized as follows: The work relevant to the research direction is described in Section II. Section III describes in detail the design and experimental implementation details of our proposed LRDE-Net, including the LRF strategy, CNI strategy, and IDE strategy. The results of the comparison and ablation experiments are thoroughly described and analyzed in Section IV. Finally, Section V concludes this article.

II. Related Work

Researchers have mostly presented algebraic operation-based approaches and image transformation-based approaches for change detection throughout the past few decades. Algebra-based methods are mainly classified into change vector analysis [9], image quantization [10], image differencing [11], image regression [12]. By computing the bitemporal images' pixel differences and then categorizing the pixels into two groups—change and unchanged—using clustering or a threshold, the approach arrives at the final results for change detection. Image transformation-based methods go for change detection by methods, such as flow cap transform (KT) [13], principal component analysis (PCA) [4]. It highlights the specific change information in the image by mapping the bitemporal image features into a new feature space. It makes use of machine learning methods like decision trees (dt) [14] and support vector machines (svm) [15] to identify the modified pixels based on the characteristics of the bitemporal pictures. High- and very-high-resolution pictures are increasingly employed in change detection applications as remote sensing image technology advances. Rich texture information can be found in high-resolution photographs. The manually designed features lack advanced semantic information, have low robustness, and cannot adapt to complex change detection scenarios.

As deep learning technology develops swiftly, more and more neural networks are being used in the field of change detection in order to get more accurate feature representations. The field of change detection has shown success using deep learning.

Long et al. [16] suggested a fully convolutional network (FCN) approach to picture segmentation. To avoid complicated connections while preserving the location data of the output characteristics and enabling pixel-level output, the approach employs fully convolutional layers rather than fully connected layers. FCNs are also becoming more and more popular in the field of change detection, and there are some similarities between the change detection issue and the image segmentation assignment. The FCN generates the change detection map by classifying the features. Villa et al. [17] proposed the full convolutional neural network-based UNet, which introduces a skip-connectivity feature fusion method and is able to obtain fine feature representations. Using a multilevel dense skip connection to collect multiscale characteristics and mitigating the pseudovariations in the detection results, Li et al. [18] built an enhanced UNet++ network. The aforementioned single-branch network, however, is unable to utilize the bitemporal images, which results in the network losing some of the characteristic information. Siamese networks have been used to identify changes in order to resolve this issue. Due to their effectiveness, siamese networks are quickly becoming commonplace.

The parameters are the same for the two Siamese branches. The Siamese network receives bitemporal images as input, and the metric module subsequently extracts the results of the change detection. The image early fusion network (FC-EF), the Siamese concatenated network (FC-Siam-conc), and the Siamese differential network (FC-Siam-diff) are three variations Daudt et al. [19] designed based on the Unet network. Through experiments, it is shown that the Siamese network has better performance, but FC-Siam-conc and FC-Siam-diff cannot obtain multiscale features well for change detection in complex scenes. Multiscale features combine the feature maps from several phases to produce information on the image location's underlying details as well as the image's higher-level semantic details. Shi et al. [20] spliced the multiscale feature maps of different stages and used a depth supervision module on the two largest scales of the feature maps, thus extracting better image features. By creating a multiscale, multilevel fusion module, Lv et al. [21] presented the CLNet model to enhance the network's capacity to extract image information. Zhao et al. [22] proposed a cross-stage feature combination network to acquire comprehensive information regarding picture features. This network involves a dual-stream encoder that retrieves features, which are subsequently inputted into a feature unification module.

To take use of the temporal-spatial correlations on the bitemporal images, the STANet [23] model is constructed, which includes a self-attention mechanism. Feng [24] et al. proposed a cross-temporal JointAttention block based on SelfAttention and CrossAttention. Fang et al. [25] introduced a novel deep learning architecture called SNU-Net, which combines a Siamese network with a dense skip-connected network to effectively acquire multiscale feature representations. Zhang et al. [26] developed spatial pyramid pools with various sized convolutions to expand the network's receptive field. Xu [27] et al. used the full attention pyramid module (FAPM) to acquire in-depth multiscale contextual information about an image using channel attention to fuse semantic features built from different channels. A network (BITNet) employing transformer was developed by Chen et al. [28], where bitemporal pictures are fed to the model and feature extraction is first carried out using a convolutional network, and then the features obtained are modeled by the constructed transformer encoder. This network allows for more detailed feature representations by extracting semantic information over long periods of time. For extracting target features at various sizes, Zhang et al. [29] presented a multiscale feature combination, multiattention Siamese Transformer network. The structure of this network is complicated, and its computational efficiency is low. Zhou et al. [30] introduced a dual cross-attention transformer (DCAT) network. This network is designed to extract both low-frequency and high-frequency information from input images through the computation of two distinct types of cross-attention features. In addition, the DCAT network demonstrates the ability to differentiate between regions that undergo changes and those that remain unchanged. DCAT's drawback is that it must manually establish the number of channels, necessitating extensive testing to get the optimal performance.

Despite the advancements made by the aforementioned approaches, they still exhibit limitations in effectively discerning pseudochanges in surface confusion. There is still some room for improvement in extracting feature representations, leading to the results of blurred change regions and sticky change detection regions. Therefore, we propose an innovative network for remote sensing images change detection and design a new strategy for expanding the sensing field and increasing the discriminative ability for pseudochange regions.

III. Proposed Method

A. Overview

In this section, the LRDE-Net's overall motivation and its architecture are presented, and the strategies that are proposed are described in detail. The four main components of the proposed LRDE-Net are the feature extractor, LRF strategy, CNI strategy, and IDE strategy, as shown in Fig. 2. The feature extractor gathers multiscale features from the bitmaporal images in order to learn representative characteristics for CD.

1) The feature extractor initially creates two feature maps $X^{(1)}, X^{(2)} \in \mathbb{R}^{C\times H\times W}$ to learn differentiating characteristics from a pair of bitemporal images covering the same area. $H \times W$ is the scale of each feature map, and C is the channel dimension of the feature vector.
2) Subsequently, $X^{(1)}, X^{(2)}$ are concatenated in the channel dimension to generate $X^{(3)}$. $X^{(3)}$ is utilised in the LRF strategy to acquire a window that is both lengthy and narrow, hence enabling the model to gather extensive global contextual information. This information is crucial for the effective functioning of the change detection network. By employing the LRF approach, it is possible to get the feature map representation denoted as X_{LRF}.
3) Simultaneously, $X^{(3)}$ is fed to the CNI strategy connected in parallel with the LRF strategy. This strategy is an efficient channel enhancement strategy that can improve the performance of various deep CNN architectures by

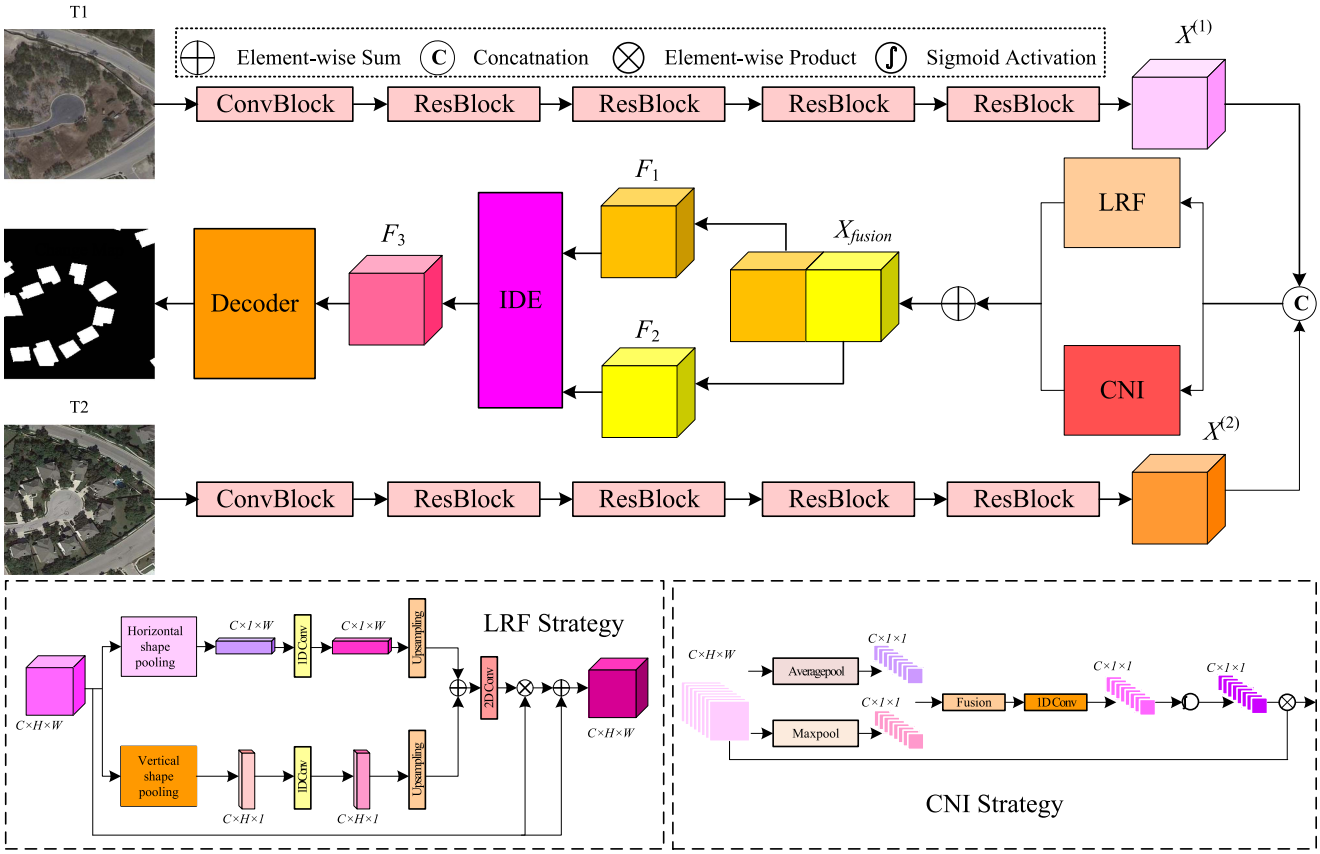

Fig. 2. Overall architecture of the proposed LRDE-Net.

generating channel interaction information through fast 1-D convolution. With the CNI strategy, we can obtain the feature map representation as X_{CNI}. Subsequently, an addition operation is executed on the features acquired via the LRF strategy and the features gained through the CNI method in order to derive the fused feature. The value of the variable can be expressed as X_{fusion}.

4) The fused feature map is partitioned into two feature maps of equal channel dimensions, which are subsequently inputted into the image differential enhancement approach to get the difference map. Subsequently, the disparity image is subjected to upsampling in order to get the predicted change map.

B. Feature Extractor

The utilization of deep convolutional neural networks in remote sensing has become prevalent due to its exceptional performance in the field of computer vision. [31]. Several studies have explored various applications in the field of remote sensing, including land use categorization [32], semantic segmentation [33], and change detection [34]. It has been demonstrated in several earlier research that using CNN feature extractors with pretraining parameters promotes model convergence [23]. Therefore, we constructed a ResNet-based feature extractor and loaded parameters from Resnet-18 [35] pretrained on Imagenet [36]. The effective implementation of CD necessitates the utilization of pixel-level prediction and is enhanced by the incorporation of dense features derived from the (FCN) approach.

To extract characteristics of various sizes and levels from bitemporal images in the encoder, we employ a fully convolutional Siamese network. The network is made up of two comparable subnets with the same weights, each of which extracts deep features from the related temporal phase image, thereby acquiring representative temporal data in the same feature region. The feature extractor is based on Resnet 18, which has four residual layers as illustrated in Fig. 2. The structure is the same for every residual layer. The latter four layers result in feature channels of 64, 128, 256, and 512, respectively. The feature maps produced by the final four layers are convolutioned to 96 to achieve a compromise between accuracy and efficiency. The final three feature maps are then convolutioned down to 1/4 of the original picture size. This method may be used to produce four feature maps with the same size and quantity of channels. Greater semantic precision and less location information are included in the higher-level characteristics of the neural network. Although the bottom-level feature map provides rich geographical information, it has less semantic information. To create improved feature representations, we combine the neural network's high-level feature information with its underlying feature information. The four above feature maps are combined

in the channel's dimension by splicing procedures before being sent into the convolution to produce additional discriminative features.

C. Large Receptive Field Strategy

Due to their superior capacity to capture high-level semantics, FCNs-based methods [25] have achieved major advancements in change detection. For a 2-D input tensor $x \in \mathbb{R}^{H\times W}$, H and W stand for the input tensor's width and height, respectively. The kernel dimension is (h, w). The output $y_{i_0,,j_0}$ after the pooling operation is a 2-D tensor. The formula for the average pooling operation can be expressed as

$$y_{i_o,j_0} = \frac{1}{h \times w} \sum_{0\le i<h} \sum_{0\le j<w} x_{i_o\times h+i, j_o\times w+j}. \tag{1}$$

The region of formula space corresponds to a pooling window with dimensions $h \times w$. The change detection tasks have been used to acquire long-range contexts by using the above-average pooling techniques [23]. But most of these methods are based on stack convolution and pooling operations. Because of the limited effective field of view, they cannot handle various different classes of complex scenarios [37], [38].

In some scenarios, the target object is a long-band structure or a discrete distribution. When using a large square pool window, it inevitably contains contaminated information from irrelevant regions [39] and thus does not solve this problem well. Self-attention or nonlocal modules [40], [41] can be used to improve the ability to construct long-range correlations. However, when computing large similarity matrices for each spatial location, it consumes a large amount of memory and takes up a lot of computational time. Expansion convolution has also been employed to model a large range of contexts. Dilated convolution expands the receptive field of the convolutional neural network without requiring the introduction of additional parameters. These methods all have the disadvantage of exploring the provided feature map throughout a rectangular window. They are less able to capture long-banded structures that are widespread in real-world scenes or that have discretely distributed contexts.

To solve this problem, we use strip pooling [42] to obtain the context information of the LRF. We designed the LRF strategy, which performs strip pooling from horizontal and vertical directions for obtaining a LRF. The structure of our proposed LRF strategy is shown in Fig. 2.

For a 2-D tensor $x \in \mathbb{R}^{H\times W}$, the pooling window for strip pooling is (H, 1) or (1, W). The designed LRF strategy performs maximum and average pooling operations on rows or columns. The output y after the horizontal pooling operation can be expressed as follows:

$$y_i^h = \text{mean}(x_{i,j}) + \max(x_{i,j}) \tag{2}$$

where $0 \le j < W$, mean($x_{i,j}$),max($x_{i,j}$)denote the average and maximum values taken for the features in row i, respectively. The output y after the vertical maximum pooling operation can be expressed as follows:

$$y_j^v = \text{mean}(x_{i,j}) + \max(x_{i,j}) \tag{3}$$

where $0 \le i < H$ denote the average and maximum values of the features for the jth column, respectively.

Increasing the network's receptive field has been demonstrated to be useful for enhancing network performance in earlier studies [43]. For an input tensor $x \in \mathbb{R}^{H\times W}$, x is fed into two parallel paths, one with a horizontal strip pooling operation and the other with a vertical strip pooling operation. We do average pooling operation and maximum pooling operation on both pathways to gain more fine-grained long-range context information. The image's context information is obtained using an average pooling operation, while the image's texture information is preserved using a maximum pooling process. The current position and its neighboring features are then modulated by a 1-D convolution with a kernel size of three using the fused feature maps on each route. Then, the horizontally and vertically pooled feature maps are upsampled and resized to $H \times W$, This produces two feature maps $y^h_{c,i,j}$and $y^v_{c,i,j}$. We fuse these two feature maps for more global contextual information.

$$y_{c,i,j} = y^h_{c,i,j} + y^v_{c,i,j}. \tag{4}$$

The output z can be obtained by the following equation:

$$\mathbf{z} = \mathbf{x} \otimes \sigma(f(\mathbf{y}))) \tag{5}$$

where σ stands for the sigmoid activation function and f represents a 1×1 convolution.

In contrast to global average pooling, which considers the entire elemental map, the LRF strategy considers long and narrow ranges without the need to establish mostly unnecessary connections between locations that are far apart. More lightweight than these attention-based modules [39], [43], they require a lot of computation to establish the relationship between each pair of locations. First, the LRF strategy employs a long kernel shape in one spatial dimension so that relationships with a large range of receptive fields can be obtained. By keeping a narrow kernel size in the other spatial dimension, it can capture local contextual information while avoiding interference from irrelevant regions. The designed LRF strategy can obtain both local and global network information.

D. Cross-Channel Interaction Strategy

We design CNI, an ECA-based [44], CNI approach, to guarantee the efficacy and efficiency of gathering local CNI information. Since the designed CNI strategy aims to capture local CNIs efficiently, it is very necessary to determine the scope of local CNIs. For convolutional blocks with different numbers of channels, the optimized coverage of the interactions can be set by manual tuning. However, manual tuning through cross-validation can be time-consuming and computationally expensive. Group convolution has been applied to convolutional neural networks to improve the performance of the network [45]. High-dimensional (low-dimensional) channels and long-range (short-range) convolution are the corresponding concepts in group convolution. In a similar vein to group convolution, the coverage of the interaction is proportional to the channel dimension. The nonlinear mapping relation between C and k is

expressed as following:

$$C = \phi(k) = 2^{(\gamma * k - b)}. \tag{6}$$

The following equation may be used to adaptively determine the kernel size for the channel dimension *C:*

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{\mathrm{odd}}. \tag{7}$$

In this article, b is set to 2 and γ to 1. $|t|_{\mathrm{odd}}$ denotes the nearest odd number to t. The process of mapping involves establishing a correspondence between high-dimensional channels and long-range convolutions, as well as between low-dimensional channels and short-range convolutions.

Fig. 2 displays the framework of our suggested CNI strategy. Using only average pooling will cause some fine-grained information to be ignored. The maximum pooling operation can obtain more salient information. Here, we use both functions to obtain richer feature information. For an input tensor $x \in \mathbb{R}^{H \times W}$, x is first subjected to global average pooling and global maximum pooling by GAP and GMP to obtain the feature maps *Fa* and *Fm*, respectively, *Fp* is generated through the addition of the two feature maps. Based on the channel dimensions of *Fp*, the adaptive convolution kernel size k is obtained by using (9). Perform 1-D convolution operation on *Fp*, and then perform sigmoid excitation function operation to obtain the weight *Mcni* of each channel number *x*.

Finally, F is multiplied with the weight *Mcni* to obtain the feature *Fcni* after CNI, where F is the feature input to the CNI strategy

$$F_{\mathrm{cni}} = M_{\mathrm{cni}} \otimes F. \tag{8}$$

We aggregate the features obtained from the LRF strategy and the CNI strategy, which are used to leverage a LRF and more discriminative channel information. We accomplish the feature fusion by performing an element-wise sum. We do not use cascading operations because they consume more computational resources. With this fusion approach, feature representation is effectively enhanced.

E. Image Difference Enhancement Strategy

The difference feature map is usually obtained using $|F_1 - F_2|$. However, in some cases, the magnitude of the difference in some change regions may not be obvious enough, which can cause the problem of omission and misdetection of change regions. To solve these issues, we provide an IDE strategy that gives the F_1 and F_2 feature maps weights so as to enhance the difference feature maps' capacity for discrimination. The structure of the IDE strategy is shown in Fig. 3. Fusing information from images taken at various periods is also essential for change detection since the anticipated change map is jointly produced by the two feature maps F_1 and F_2. We first fuse the features of F_1 and F_2 by adding them element-wise

$$F = F_1 + F_2 \tag{9}$$

where F_1 and F_2 denote the feature maps at *T1* and *T2* moments, respectively, and F denotes the fused. After that, run a global

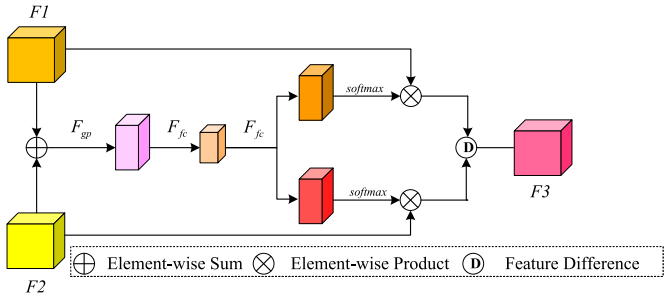

Fig. 3. Illustration of the proposed IDE strategy.

average pooling operation to produce global information. The following is the precise calculation formula for the c th element of *s:*

$$f_c = \frac{1}{W \times H} \sum_{j=1}^{W} \sum_{k=1}^{H} X_{\mathrm{fuse}}^{c}(j,k). \tag{10}$$

A fully connected layer generates a compact feature $z \in R^{d \times 1}$

$$z = \mathcal{F}_{fc}(\mathbf{s}) = \delta(\mathcal{B}(W\mathbf{s})) \tag{11}$$

where δ means the ReLU function, $\mathcal{B}$ stands for batch normalization, and $W \in \mathbb{R}^{d \times c}$.

We weighted the feature maps at *T1* and *T2* for the purpose to improve the differentiation capability of creating differential feature maps. First, we generate confidence weights for the feature maps at each time using fully connected and softmax activation functions, and then each feature map is multiplied by the corresponding confidence weight. Finally, the temporal feature maps endowed with weights are subjected to the difference operation. The formula is shown as follows:

$$w_1 = \mathrm{softmax}\left(\mathcal{F}_{fc}^{1}(p)\right), \quad w_2 = \mathrm{softmax}\left(\mathcal{F}_{fc}^{2}(p)\right) \tag{12}$$

$$F_3 = w_1 F_1 - w_2 F_2 \tag{13}$$

where w_1,w_2 denote the obtained confidence weights and F_3 denotes the obtained difference feature map.

F. Loss Function

Following the IDE strategy, the feature map is upsampled to the scale of the input image by running it through two 3×3 deconvolution layers. While H_0 and W_0 stand for the height and breadth of the original input image, respectively, $D \in R^{H_0 \times W_0}$ stands for the updated feature map. We employ a contrast loss function during the training stage to optimize the network. The change map is created during the testing process by establishing a threshold that has been set. The split of the number of change samples and unchanged samples in the change detection task is quite imbalanced. Only a small percentage of all pixels are made up by the change pixel, which might bias the network while training. Batch-balanced contrast loss (BCL) [23] is used to decrease the effects of category imbalance. With the suggested LRDE-Net, we can determine the distance map, where *B* stands for the batch of samples, and H_0 and W_0 represent the height and width of the supplied image, respectively. The loss function's

formula is shown as follows:

$$L\left(D^*, M^*\right) = \frac{1}{2}\frac{1}{n_u}\sum_{b,i,j}\left(1 - M^*_{b,i,j}\right) D^*_{b,i,j} + \frac{1}{2}\frac{1}{n_c}\sum_{b,i,j} M^*_{b,i,j}\,\mathrm{Max}\left(0, m - D^*_{b,i,j}\right) \quad (14)$$

where n_u,n_c denote the amount of shifting and unchanging pixel pairings, respectively. *b,i*, and *j* denote the batch, width and height, respectively. $M^*_{b,i,j}$ denote the pixel values corresponding to the binary labeled map. Changes to pixel pairs with parameterized distances greater than *m* have no effect on the loss function's result. *m* in this study is 2.

IV. Experimental Results and Analysis

A. Dataset Description

1) WHU-CD [2]: It was created using data from remote sensing photographs taken both prior to and following earthquakes in the New Zealand city of Christchurch. It has a couple of images with a combined resolution of 0.2 m and 32 507×15 354 pixels. The pair consists of one image captured in April 2012, which encompasses a total of 12 796 buildings, and two subsequent photographs obtained in 2016, which collectively encompass 16 077 structures. It is located inside the Christchurch region of New Zealand and spans 20.5 square kilometres. During this period, many buildings were rebuilt or newly constructed. The images were cropped to 256×256 pixels based on the principle of nonoverlapping. Training, validation, and testing samples are distributed in an 8:1:1 ratio.

2) LEVIR-CD [23]: This dataset is a huge, openly accessible dataset that includes 637 pairs images with a size of 1024×1024 in total. These photos were collected from more than 20 distinct locations spread throughout many Texas cities. The images represented in the collection span a period of time ranging from 5 to 14 years. The dataset contains small changes, such as changes in the increase and decrease of buildings. The dataset contains many types of buildings, like residential buildings, cottages, garages, etc. The image is cropped and rotated to generate an image of 256×256 pixels. According to the prescribed dataset segmentation protocol, the distribution of data is divided into three sets, namely the training set, validation set, and test set, with a ratio of 7:2:1.

3) CDD [46]: The data utilized in this study is obtained from remote sensing photographs of the same geographical region captured by Google Earth, showcasing seasonal variations. The dataset has seven pairs of photos, each with a size of 4725×2700 pixels, as well as four pairs of images with a resolution of 1900×1000 pixels. The picture resolution spans from 0.03 to 1 metre per pixel. The dataset encompasses a diverse range of items with varying dimensions, including automobiles, big structures, roadways, etc. The photos undergo cropping and rotation processes in order to produce images with dimensions of 256×256 pixels. The CDD dataset has an overall total of 16,000 picture pairings, with 10,000 pairs allocated for training, 3000 pairs for validation, and the remaining 3000 pairs for testing purposes.

B. Evaluation Metrics

To assess the effectiveness of the suggested algorithm, we conducted a quantitative evaluation of the proposed network. This evaluation involved measuring precision, recall, F1 score, and overall accuracy (OA). Precision refers to the proportion of accurately identified altered pixels in relation to the overall number of pixels identified as altered. Recall refers to the ratio of accurately identified altered pixels to the total number of altered pixel samples inside the label. The F1 score is a complete assessment metric that reflects the performance of the suggested algorithm, as it is determined by precision and recall. The term "OA" refers to the proportion of accurately classified pixels across all samples. The calculating formulas are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (15)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (16)$$

$$\text{F1} = \frac{2}{\text{Precision}^{-1}+\text{Recall}^{-1}} \quad (17)$$

$$\text{OA} = \frac{\text{TP+TN}}{\text{TP} + \text{FP+TN+FN}}. \quad (18)$$

The quantity of pixel samples that are successfully identified as change pixel samples is denoted by TP, whereas the quantity of pixel samples that are misidentified as change pixel samples is denoted by FP. The quantity of pixel samples that are accurately identified as nonchange pixel samples is indicated by the TN, whereas the quantity of pixel samples that are misidentified is indicated by the FN.

C. Implementation Details

The NVIDIA Titan RTX (24 GB) GPU running the PyTorch architecture is utilized to carry out the LRDE-Net algorithm suggested in this work. The network model parameters are optimized at the time of training using adam [47]. The dataset was configured with a batch size of 16, and the learning rate was set at 10^{-4}. The epoch value of 200 was used for the dataset in order to achieve model convergence. Following each training session, the trained model undergoes testing on the validation set. The best-trained model is selected for evaluating the performance on the test set.

On the WHU-CD, LEVIR-CD, and CDD datasets, we compared our proposed algorithm to SOTA approaches. All comparison algorithms were reproduced using the same dataset. The comparison algorithms use the parameter settings from the original paper. All experimental procedures were conducted using identical hardware and software configurations. The LRDE-Net is evaluated against eight other state-of-the-art change detection methods on three datasets, namely FC-EF [19], FC-Siam-conc [19], FC-Siam-diff [19], STANet [23], BIT-CD [28], SNUNet/48 [25], ChangeFormer [48], DMINET [24].

D. Results and Analysis on WHU-CD

1) Quantitative Analysis on WHU-CD: Table I displays the precision, recall, F1, and OA of each method on the WHU-CD

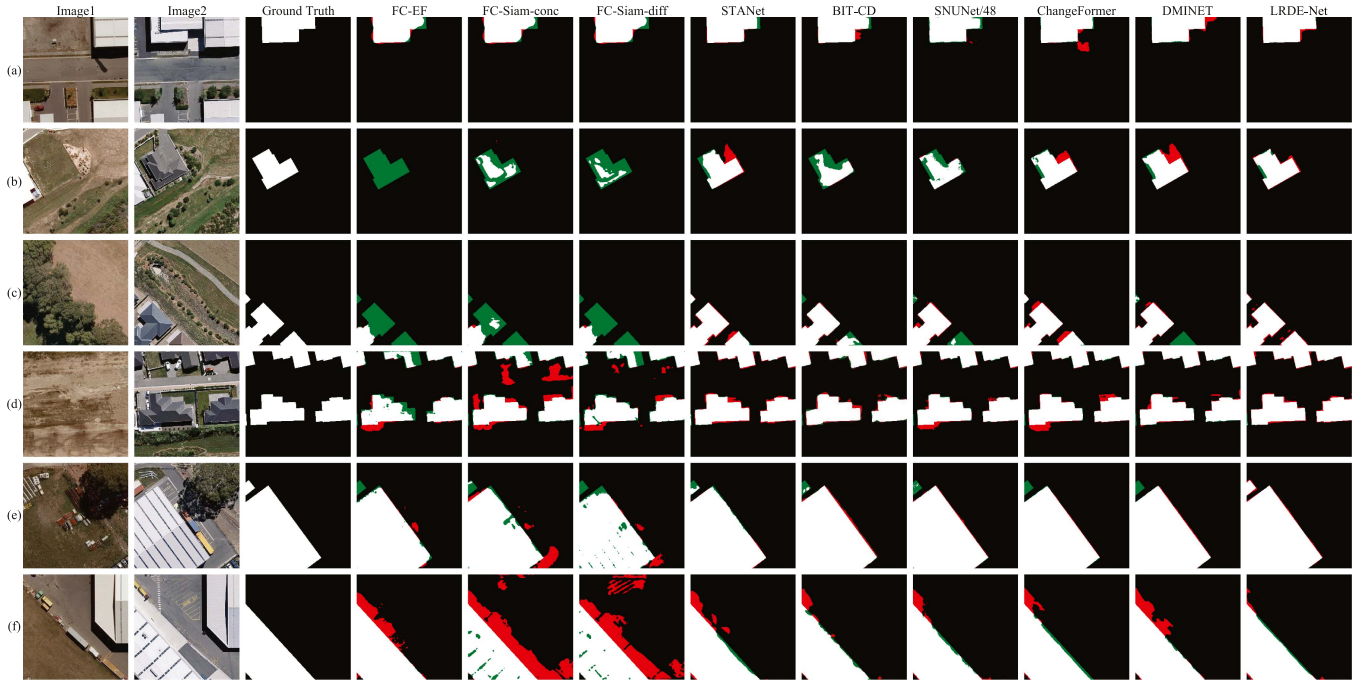

Fig. 4. Visualization comparison results of different methods on WHU-CD dataset. (a)–(f) Denote six different samples in the dataset.

TABLE I
COMPARISON RESULTS ON WHU-CD DATASET

Model	Precision	Recall	F1	OA
FC-EF	82.28	70.66	76.03	97.92
FC-Siam-conc	40.09	73.84	51.97	93.63
FC-Siam-diff	38.82	71.80	50.40	93.40
STANet	91.25	86.18	88.64	98.97
BIT-CD	86.07	85.61	85.84	98.68
SNUNet	88.35	87.80	88.07	98.89
ChangeFormer	**93.44**	85.70	89.40	99.12
DMINET	85.75	88.02	86.87	98.76
LRDE-Net	91.57	**90.10**	**90.83**	**99.15**

dataset. We have bolded the best performing values. The table shows that, when compared to other approaches, LRDE-Net has the best recall, F1, and OA. Among these eight algorithms, FC-EF has the lowest precision, recall, F1, and OA. In addition, the ChangeFormer algorithm ranks second among these eight algorithms with the highest precision, but its recall is relatively low. Compared to the ChangeFormer algorithm, our LRDE-Net improves 4.4%, 1.43%, and 0.03% on recall,F1, and OA, respectively. Among the eight algorithms considered, STANet is ranked third. In addition, our LRDE-Net demonstrates superior performance compared to STANet with respect to precision, recall, F1, and OA. The LRDE-Net has a superiority of 3.92% and 2.19% above the STANet with respect to F1 and OA, respectively.

2) Visualization Analysis on WHU-CD: We use the colors black, white, red, and green to denote the sections that correlate to TP, TN, FP, and FN in the visualization comparison section. For visualizing the impact of each network model's change detection on the WHU-CD dataset, we randomly choose a few images. Fig. 4(a) and (b) contains obvious radial changes with severe interference from concrete or asphalt-covered surfaces. The FC-EF, FC-Siam-diff, FC-Siam-conc, and BIT algorithms contain more missed regions. STANet, ChangeFormer exist to recognize a part of the road surface as a building. Fig. 4(c) contains tiny buildings, and the FC-EF, FC-Siam-diff, FC-Siam-conc, BIT, and SNUNet networks are unable to detect or detect a very small part of the tiny buildings. Compared to the LRDE-Net algorithm, ChangeFormer has more false detection regions. Fig. 4(d) shows dense buildings containing seasonal changes, interfered with by similarly colored concrete or asphalt-covered surfaces. Except for the LRDE-Net network, all other networks have serious false detection situations for cement or asphalt-covered surfaces. Fig. 4(e) and (f) contains changes in atmospheric conditions and large and small buildings. The FC-Siam-diff and FC-Siam-conc networks exhibit erroneous identification of some road expansions as alterations in building structures. Except for the LRDE-Net network, all other networks have serious omissions for small buildings. In the aforementioned pictures, all of our LRDE-Net networks can clearly discriminate between changing and unchanging regions using the proposed IDE method, and they are also able to avoid interference from pseudochanges and radial changes.

E. Results and Analysis on LEVIR-CD

1) Quantitative Analysis on LEVIR-CD: We performed tests on the LEVIR-CD dataset to assess the performance of the proposed approach, and the results are displayed in Table II. LRDE-Net has the highest recall (93.17%), F1(90.78%), OA(99.09%).The ChangeFormer algorithm has the maximum precision, but its recall is rather poor, which will lead to more issues with omission detection. Our proposed

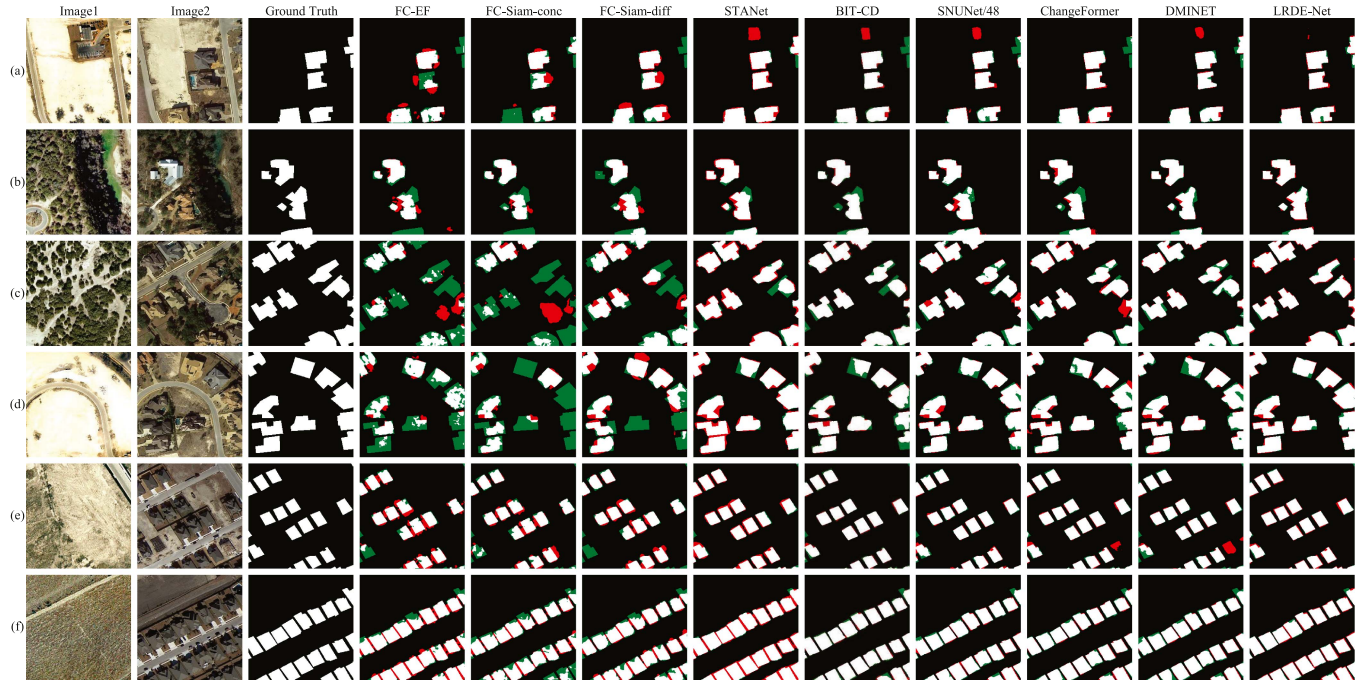

Fig. 5. Visualization comparison results of different methods on LEVIR-CD dataset. (a)–(f) Denote six different samples in the dataset.

TABLE II
COMPARISON RESULTS ON LEVIR-CD DATASET

Model	Precision	Recall	F1	OA
FC-EF	82.27	66.28	73.41	97.55
FC-Siam-conc	86.81	67.66	76.05	97.83
FC-Siam-diff	86.55	74.38	80.00	98.11
STANet	80.99	91.21	85.79	98.46
BIT-CD	91.95	88.57	90.23	99.02
SNUNet	91.66	88.48	90.04	99.00
ChangeFormer	91.53	88.86	90.17	99.01
DMINET	**92.53**	87.32	89.85	98.99
LRDE-Net	88.50	**93.17**	**90.78**	**99.04**

TABLE III
COMPARISON RESULTS ON CDD DATASET

Model	Precision	Recall	F1	OA
FC-EF	76.56	43.49	55.47	91.76
FC-Siam-conc	88.00	53.58	66.61	93.66
FC-Siam-diff	88.49	51.53	65.14	93.49
STANet	88.97	94.31	91.56	97.95
BIT-CD	95.86	94.59	95.22	98.88
SNUNet	**96.82**	96.72	96.77	**99.24**
ChangeFormer	95.28	93.83	94.55	98.72
DMINET	95.78	95.88	95.83	99.02
LRDE-Net	96.54	**97.08**	**96.81**	99.21

LRDE-Net algorithm can better detect those changing pixels that are not easy to detect. The ChangeFormer algorithm is ranked second among the eight algorithms. On recall, F1, and OA, our LRDE-Net performs better than the ChangeFormer method by 4.31%, 0.61%, and 0.03%, respectively. According to these statistical results, the proposed LRDE-Net algorithm performs better than other comparable methods in terms of detection performance on the LEVIR-CD dataset.

2) Visualization Analysis on LEVIR-CD: We randomly selected some images for visualizing the change detection effect of each network model on the LEVIR-CD dataset. Fig. 5(a) and (d) shows several structures that are occlusive. The change detection produced by LRDE-Net has the fewest missed detections, as can be observed from the detection results. The detection results show that the change detection produced by LRDE-Net has the fewest missed detections. Fig. 5(b) is disturbed by obvious light changes and also contains some occlusions that make it difficult to distinguish the changing regions. With the exception of the LRDE-Net approach, none of the other methods are capable of detecting the change area located in the center of the picture. The LRDE-Net stands out as it produces the most accurate change maps with labels. Fig. 5(c) contains changes in interference from occlusion and some added roads. Missed detections are prevalent in FC-EF, FC-Siam-diff, and FC-Siam-conc. Our approach has the fewest false detection regions when compared to STANet, BIT, SNUNet, and ChangeFormer. Dense structures may be seen in Fig. 5(e) and (f), which are disrupted by surfaces covered in concrete or asphalt. Due to their ineffective feature extraction capabilities, FC-EF, FC-siam-diff, and FC-siam-conc have poor detection. LRDE-Net has the fewest false detections and missed detection regions compared to other approaches, and it can display more regular and full roofs, proving that the proposed LRDE-Net method is more resilient.

F. Results and Analysis on CDD

1) Quantitative Analysis on CDD Dataset: The assessment metrics for change detection for each approach on the CDD dataset are shown in Table III. The precision and F1 of FC-Siam-diff are 11.93% and 9.67% higher than those of FC-EF, respectively. The chart shows that all assessment metrics for

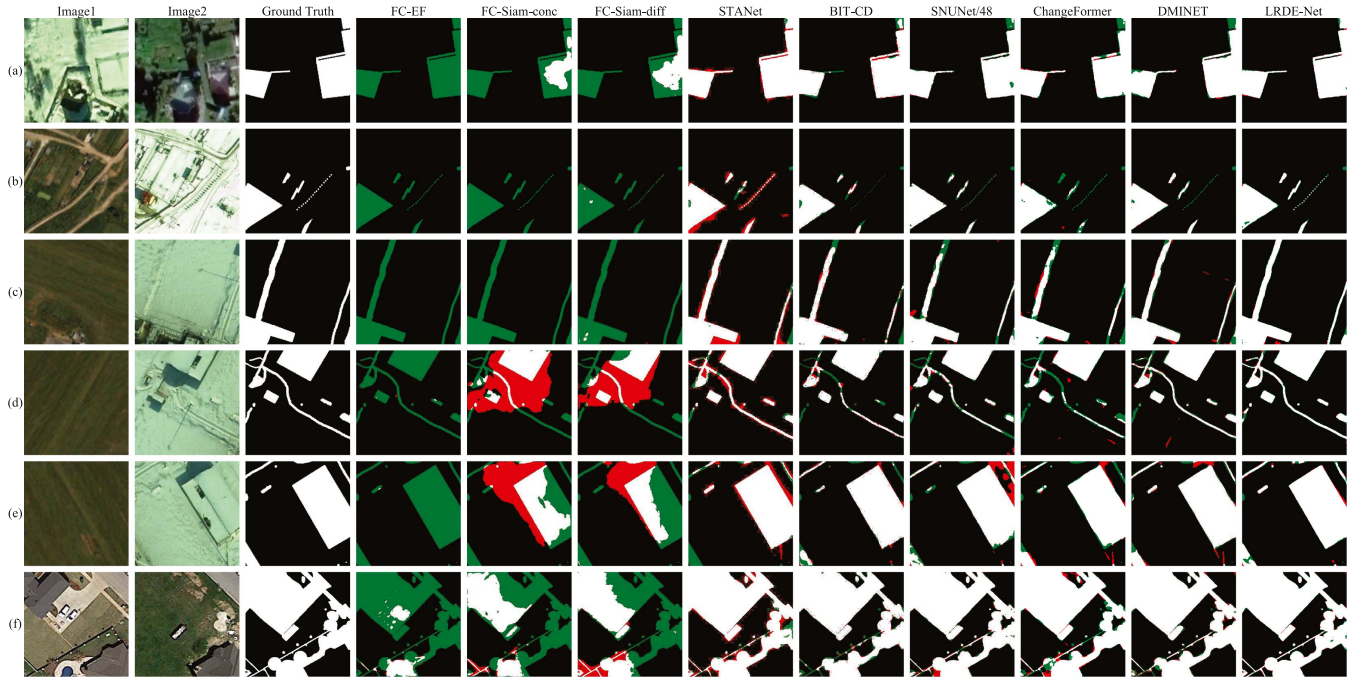

Fig. 6. Visualization comparison results of different methods on CDD dataset. (a)–(f) Denote six different samples in the dataset.

FC-EF are the lowest. It demonstrates that the Siamese network outperforms the unilateral network in terms of increasing network performance. The LRDE-Net method outperforms the other methods and has the highest F1 value. It is 0.04% and 0.36% higher than SNUNet, which is ranked second in F1 and recall, respectively. The LRDE-Net method is 1.26%, 3.25%, 2.26%, and 0.49% higher than Changeformer, which is ranked third in precision, recall, F1, and OA, respectively. The aforementioned quantitative results demonstrate the advantages and efficacy of the LRDE-Net approach suggested in this study.

2) Visualization Analysis on CDD: Since some of the items in the CDD dataset are sparse and tiny, detecting changes is made more challenging. Fig. 6(a), (c), and (d) contains changes in roads and buildings. Fig. 6(b) contains dense, small target changes disturbed by seasonal changes. Fig. 6(e) contains changes in large buildings and roads, disturbed by snow cover. Fig. 6(f) contains dense building variations of different sizes. The FC-EF, FC-Siam-diff, and FC-Siam-conc approaches can only identify a small fraction of the changing objects, as shown in Fig. 6, and suffer from a significant number of false detections and missed detection issues. STANet uses spatio-temporal attention to improve the network model's capacity to extract its features; however, it is insensitive to the target object's edge information and makes a lot of false detections. Because global information is available, using Transformer, BIT, and Changeformer can keep the integrity of building changes at the border. However, because local complementary information is not available, fine-grained changes cannot be found. SNUNet obtains a more adequate feature representation through multiscale convolutional feature fusion, but due to the limited receptive field, some missing target objects appear during detection. In the prediction of long-shaped buildings and roads, our LRDE-Net predictions are closest to the real labels due to the fact that our proposed LRF strategy can integrate both global and local information, which has better feature representation capability. The change regions predicted by LRDE-Net are complete and accurate, which can suppress the effects caused by seasonal changes.

TABLE IV
ABLATION EXPERIMENTS RESULTS ON WHU-CD DATASET

Method	Precision	Recall	F1	OA
baseline	86.38	86.36	86.37	98.73
w/o SP	89.12	87.97	88.54	98.83
w/o eca	90.39	88.96	89.67	99.01
w/o adm	90.45	89.87	90.16	99.04
LRDE-Net	**91.57**	**90.10**	**90.83**	**99.15**

G. Ablation Analysis

To assess the efficacy of individual strategies employed in the LRDE-Net algorithm, ablation experiments were conducted using the WHU-CD dataset as an example. Table IV shows the performance of the network under different experimental conditions. The abbreviation "w/o" denotes the elimination of a certain strategy. From Table IV, it can be seen that Without the LRF strategy, the F1 performance drops even more by 2.29%. This is because strip pooling uses a lengthy kernel shape in a single spatial dimension to acquire a wide variety of associations in isolated regions. Meanwhile, keeping a narrow kernel size in the other spatial dimension allows for capturing local context information while avoiding interference from irrelevant regions. A change detection network integrated with strip pooling can obtain both local and global network information. As a result, the LRDE-Net algorithm's performance may be greatly enhanced.

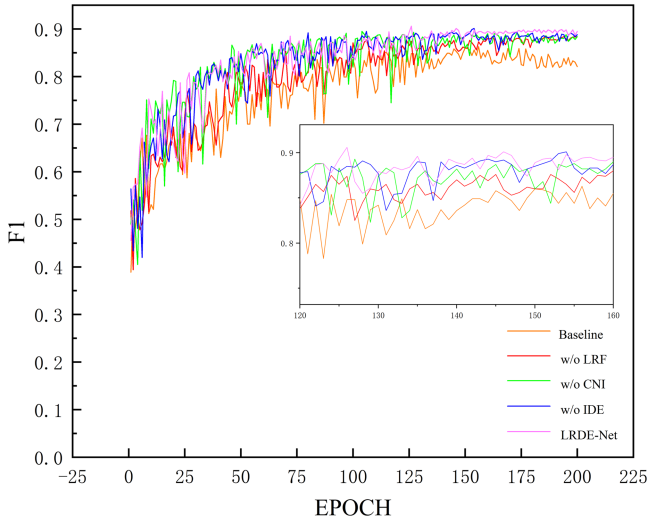

Fig. 7. Line chart of ablation experiments on WHU-CD dataset, with small windows showing localized zoomed-in details.

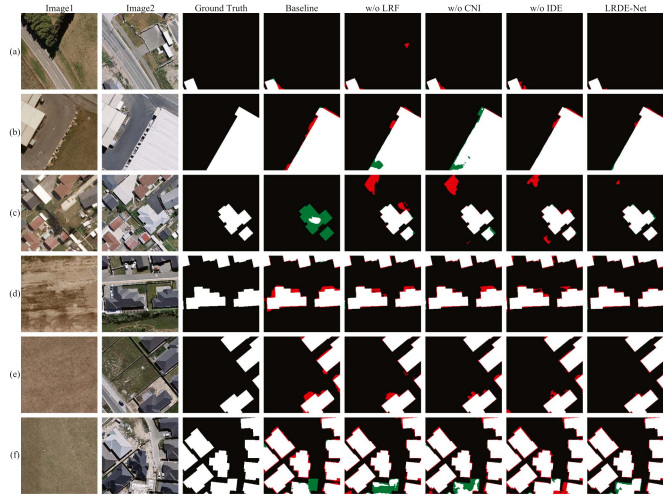

Fig. 8. Visualization comparison results of of ablation experiments on WHU-CD dataset. (a)–(f) Denote six different samples in the dataset.

It may be discovered that the F1 performance reduces by 1.16% after deleting the CNI method by looking at the performance of the LRDE-Net algorithm. The CNI strategy obtains the relationship between the feature map channels through CNIs, and it can fully utilize the remote context information and the more discriminative channel information to obtain a better representation of features. The utilization of the IDE strategy involves the allocation of weights to F1 and F2 feature maps. This approach aims to improve the discriminative capability of producing distinct feature maps and ultimately provide superior change detection outcomes. When the IDE strategy is removed, the network performance decreases by 0.67%, demonstrating the effectiveness of our proposed strategy.

Fig. 7 displays a line chart depicting ablation trials. The horizontal axis represents the epoch, while the vertical axis represents the F1 performance. The line chart consists of five separate ablation experiments, each represented by a line of a distinct color. For a clearer view, we zoom in on a segment of the performance folds during training, as shown in the small window. The graphic illustrates a progressive leveling down of performance folds with an increase in training periods. Baseline has the lowest performance, we propose that the complete network has the best performance, and when any one of the strategies is removed, the F1 performance decreases.

Fig. 8 displays the ablation experiment's visualization results. After removing the LRF strategy, due to the decrease in the range of the network receptive field, the detection results will deteriorate in the long strip-shaped changing area. When the IDE strategy is removed, the network will have more misdetections and missed detections. As observed in Fig. 8, the LRDE-Net method outperforms the other algorithms and comes the closest to the label in terms of detection performance. When removing any of the modules, there is more leakage and misdetection. The efficacy of the strategies we proposed was validated by means of ablation visualization.

V. Discussion

In the realm of remote sensing, change detection is an important and popular study path. Our proposed method outperforms existing comparison models after extensive testing. Our proposed network has good detection results on changing regions such as huge buildings, small buildings, densely dispersed buildings, and other tiny targets, as shown in Figs. 4(b) and (e), 5(f), and 6(b). Because of their small receptive fields, traditional convolutional neural networks have a limited capacity to retrieve contextual information. When using self-attention to extract visual characteristics, local detail information is lost while acquiring long-range contextual information. The LRF strategy is proposed to acquire long-range contextual information in one dimension and local information in another. Avoiding the loss of local detail information, while acquiring long-term context information. Furthermore, the proposed CNI strategy models the interaction between channels in order to generate an effective feature representation. The IDE strategy can help distinguish between change and invariant areas and improve the network's capacity to discern between them.

Following our thorough consideration, there are three different areas in change detection that might be researched in order to produce improved change detection algorithms.

1) Data augmentation for remote sensing images. For training on tasks involving change detection, numerous training samples are needed. The training samples are occasionally small and challenging to collect in real-world situations. Due to the limited number of training data, the network will have low generalization ability and can only have a good detection effect on comparable scenes or similar change items. A category imbalance and unfavorable change detection results will also occur from realistic scenes' slow area changes and minimal number of changing pixels. The generalization of the network can be improved, and the effect of category imbalance on network performance can be lessened, by researching

ways to increase the number of samples and the proportion of change pixels.

2) Design novel loss functions. The majority of the existing loss functions employed in the change detection field are designed to address the sample imbalance issue. A challenging issue for change detection is object edge detection, and there is currently only a small amount of research being done on the top of the loss function to address this issue. To enhance the effectiveness of change detection, loss functions that can resolve the edge detection problem can be investigated.
3) Study semisupervised learning methods. Numerous remote sensing images are now available because of advancements in technology. Large-scale remote sensing picture files containing change zones must be labeled manually, which is time-consuming and expensive. Semisupervised learning techniques that label a limited number of examples can solve this issue. The semisupervised learning methods that currently exist have complex frameworks, and their performance needs to be improved. Therefore, further research can be done on semisupervised learning methods with simple structure, high generality, and good performance.

VI. Conclusion

We propose the LRDE-Net model for change detection in this study. We designed the LRF strategy to increase the network's receptive field. The performance of the LRDE-Net algorithm can be greatly enhanced by expanding the network's receptive field by incorporating local and global network information. We designed the CNI strategy to efficiently obtain relationship information between feature map channels. We propose the IDE strategy for distributing weights to F1 and F2 feature maps in order to improve the discriminative capability of producing different feature maps and to get improved change detection results. The effectiveness of the proposed strategy is assessed by extensive testing on three open datasets. When compared to eight SOTA algorithms, LRDE-Net achieves the best overall performance on the datasets. Ablation tests are used to confirm the efficiency of the suggested strategy. Our algorithm can well resist the pseudovariation interference caused by light and seasons and has good robustness. Training sample is sometimes difficult to get and insufficient in circumstances for actual change detection scenarios. In further research, we will explore a semisupervised learning method for the proposed network and investigate how to boost LRDE-Net's performance with only a few samples.

References

[1] Q. Li et al., "Unsupervised hyperspectral image change detection via deep learning self-generated credible labels," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9012–9024, 2021.

[2] 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.

[3] B. Parmentier, "Characterization of land transitions patterns from multivariate time series using seasonal trend analysis and principal component analysis," *Remote Sens.*, vol. 6, no. 12, pp. 12639–12665, 2014.

[4] 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.

[5] T. Lei et al., "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[6] R. Wang et al., "Lightweight convolutional neural network for bitemporal SAR image change detection," *J. Appl. Remote Sens.*, vol. 14, no. 3, pp. 036501–036501, 2020.

[7] S. Xiang, M. Wang, X. Jiang, G. Xie, Z. Zhang, and P. Tang, "Dual-task semantic change detection for remote sensing images using the generative change field module," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3336.

[8] X. Xu, J. Li, and Z. Chen, "TCIANet: Transformer-based context information aggregation network for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1951–1971, 2023.

[9] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[10] X. Dai and S. Khorram, "Quantification of the impact of misregistration on the accuracy of remotely sensed change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Proc. Remote Sens.-A Sci. Vis. Sustain. Devlop.*, 1997, pp. 1763–1765.

[11] N. Quarmby and J. Cushnie, "Monitoring urban land cover changes at the urban fringe from spot HRV imagery in south-east England," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 953–963, 1989.

[12] P. R. Coppin and M. E. Bauer, "Digital change detection in forest ecosystems with remote sensing imagery," *Remote Sens. Rev.*, vol. 13, no. 3/4, pp. 207–234, 1996.

[13] S. Jin and S. A. Sader, "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances," *Remote Sens. Environ.*, vol. 94, no. 3, pp. 364–372, 2005.

[14] J. Im and J. R. Jensen, "A change detection model based on neighborhood correlation image analysis and decision tree classification," *Remote Sens. Environ.*, vol. 99, no. 3, pp. 326–340, 2005.

[15] M. Davy, F. Desobry, A. Gretton, and C. Doncarli, "An online support vector machine for abnormal events detection," *Signal Process.*, vol. 86, no. 8, pp. 2009–2025, 2006.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[17] M. Villa, G. Dardenne, M. Nasan, H. Letissier, C. Hamitouche, and E. Stindel, "FCN-based approach for the automatic segmentation of bone surfaces in ultrasound images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, pp. 1707–1716, 2018.

[18] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[19] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[20] 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, pp. 1–16, 2022.

[21] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[22] Y. Zhao, P. Chen, Z. Chen, Y. Bai, Z. Zhao, and X. Yang, "A triple-stream network with cross-stage feature fusion for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

[23] 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, 2020, Art. no. 1662.

[24] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[25] 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, pp. 1–5, 2022.

[26] C. Zhang, S. Wei, S. Ji, and M. Lu, "Detecting large-scale urban land cover changes from very high resolution remote sensing images using cnn-based classification," *ISPRS Int. J. Geo- Inf.*, vol. 8, no. 4, p. 189, 2019.

[27] X. Xu, Z. Yang, and J. Li, "AMCA: Attention-guided multi-scale context aggregation network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-19, 2023.
[28] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[29] H. Li, X. Liu, H. Li, Z. Dong, and X. Xiao, "MDFENet: A multiscale difference feature enhancement network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3104–3115, 2023.
[30] Y. Zhou, C. Huo, J. Zhu, L. Huo, and C. Pan, "DCAT: Dual cross-attention-based transformer for change detection," *Remote Sens.*, vol. 15, no. 9, 2023, Art. no. 2395.
[31] Y. Liang, C. Zhang, and M. Han, "RaSRNet: An end-to-end relation-aware semantic reasoning network for change detection in optical remote sensing images," *IEEE Trans. Instrum. Meas.*, early access, Feb. 22, 2023, doi: 10.1109/TIM.2023.3243680.
[32] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*.
[33] H. Chen, T. Shi, Z. Xia, D. Liu, X. Wu, and Z. Shi, "Learning to segment objects of various sizes in VHR aerial images," in *Proc. 13th Conf. Image Graph. Technol. Appl.*, 2018, pp. 330–340.
[34] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
[38] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[39] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7519–7528.
[40] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.
[41] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6656–6664.
[42] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4003–4012.
[43] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[44] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.
[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.
[46] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, 2018.
[47] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[48] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

Lele Li received the master's degree in information and communication engineering from the School of Communications and Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an, China, in 2019. He is currently working toward the Ph.D. degree in computer science and technology with the School of Computer Science and Technology, Xinjiang University, Ürümqi, China.

His research interests include computer vision and remote sensing image change detection.

Liejun Wang received the Ph.D. degree in information and communication engineering from the School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an, China, in 2012.

He is currently a Professor with the School of Computer Science and Technology, Xinjiang University, Ürümqi, China. His research interests include wireless sensor networks, computer vision, and natural language processing.

Anyu Du received the master's degree in information and communication engineering from the School of Computer Science and Technology, Xinjiang University, Xinjiang, China, in 2016.

She currently works with the School of Computer Science and Technology Xinjiang University. Her research interests include computer vision and wireless sensor networks.

Yongming Li received the master's degree in computer technology from the School of Computer Science and Technology, Xinjiang University, Xinjiang, China, in 2007.

He is currently an Associate Professor and works with the School of Computer Science and Technology, Xinjiang University. His research interests include computer vision, natural language processing, and wireless sensor networks.