R&D-Net: Integration of Registration-Net and Detection-Net for Identifying Building Changes in High Spatial-Resolution Remote Sensing Images

Zhong Chen, Tong Zheng, Junsong Leng, Jiahao Zhang, He Deng, Xiaofei Mi, and Jian Yang

***Abstract*—In the detection of building changes in high spatial-resolution remote sensing images, both the distribution position and surface characteristics of the same objects are probably different under different imaging phases, which potentially causes high false positives. In order to improve the detection accuracy of building changes, an integration network, named R&D net is proposed in this article, which comprises a registration network (R-net) followed by a change detection network (D-net). In R-net, two different phase images are accepted as inputs, corner points and their descriptors are generated to spatially align those images. After that, the spatially aligned images are fed into the D-net, and building images are detected accordingly. In this article, a multiview automatic labeling method is proposed to obtain labeling corner points. A new dataset containing 5104 image pairs is established. Experimental results demonstrate that the R-net can extract robust invariant features, and then improve registration accuracy under circumstances with obvious changes of surface feature, which is a base of D-net. Uniting pyramid pooling structure with a focal loss function in D-net, both leaky and wrong segmentations can be dramatically improved under complex scenes with many interferences. When compared with baseline methods on different high-resolution remote sensing scenes, the proposed method achieves better performance and more accurate detection results of building changes.**

***Index Terms*—Building changes, detection network (D-net), registration, registration network (R-net), segmentation.**

I. Introduction

USING multitemporal high spatial resolution remote sensing imagery for building change detection is a fundamental task in the remote sensing field that is both meaningful and challenging. Building change detection has significance in application fields of remote sensing images, such as urban planning [1], post-disaster reconstruction [2], military information monitoring [3], and geographical information establishment [4]. Usually, the input of building change detection includes two remote sensing images acquired at different temporal phases.

An obvious problem is that it is impossible to obtain spatially consistent images from different imaging periods (i.e., with inconsistent image sizes, resolutions, and geographic coordinates) due to potential differences in imaging sensors or varying imaging angles of the same satellite. Indeed, it means that the position and angle of the same building in different images are likely to be significantly different. Without image registration, even if the same building is detected in different images, it would be difficult to observe its changes. Registering image pairs before performing change detection on buildings can effectively enhance the performance of transformation detection and improve visualization effects.

Another problem that needs to be addressed is the segmentation of buildings. If the time interval between the acquisition of two images is slightly longer, surface changes (such as building demolition and reconstruction, vegetation changes, and shadow effects) will become more significant, especially for high-resolution images. This may increase the difficulty of building change detection, leading to low accuracy in building change detection and inaccurate boundary segmentation. Pixel-level building change detection is easily affected by many factors, such as mistaking large vehicles, roads, and squares for buildings and missing detection due to different imaging angles of buildings. Therefore, improving the robustness of the model is crucial.

In this article, we propose a novel change detection network (named as R&D-net) to solve registration and building change detection problems. R&D-net consists of a registration network (R-net) followed by a change detection network (D-net). The contributions of this article can be summarized as follows.

1) To improve the registration accuracy in the case of significant changes in terrain features, R-net is proposed in this article. R-net consists of two output branches: corner detection and feature vector description. To mitigate the issue of sample imbalance, focal loss is used in corner detection. To generate a high-resolution remote sensing image registration dataset, this article proposes an automatic labeling method based on the bagging theory, which can annotate corner points in remote sensing images and

Manuscript received 21 July 2023; revised 20 November 2023; accepted 17 December 2023. Date of publication 20 December 2023; date of current version 10 January 2024. The work was supported in part by the National Natural Science Foundation of China under Grant 62071456 and in part by The Major Project of High Resolution Earth Observation System under Grant 30-Y60B01-9003-22/23. *(Corresponding author: Junsong Leng.)*

Zhong Chen and Junsong Leng are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430070, China (e-mail: henpacked@hust.edu.cn; hzauljs@163.com).

Tong Zheng is with the Huawei Technologies Company Ltd., Chengdu 611730, China (e-mail: zhengtong88@hust.edu.cn).

Jiahao Zhang is with the Hikvision Digital Technology Company Ltd., Hangzhou 310051, China (e-mail: zhangjiahao19@hikvision.com).

He Deng is with the Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: denghe@wust.edu.cn).

Xiaofei Mi and Jian Yang are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100080, China (e-mail: mixf@aircas.ac.cn; yangjian@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3344939

generate effective registration dataset without the need for manual supervision.

2) To achieve precise segmentation of buildings, D-net is proposed in this article. The RebuildBlock module is used to enhance the segmentation accuracy of building profile. To address the incomplete segmentation of large-scale building change areas and the omission of small buildings, the ASPP module is employed to improve the similarity of segmentation regions. To solve the problem of the algorithm mistakenly identifying newly built roads as reconstructed buildings, this paper utilizes the nonlocal module to enable the network to obtain global image information, thereby avoiding misjudgments.

II. Related Work

A. Registration Methods

The registration methods of remote sensing images are roughly divided into gray-level information [5], [6], frequency domain [7], feature [4], [8], [9], and deep learning algorithms [10]. The first three approaches are traditional registration methods, but with low accuracies and robustness for images which have obvious features changes of ground objects [3]. Recently, convolutional neural networks (CNNs) are becoming the mainstream for image registration owing to the strong nonlinear learning capabilities [11]. The CNNs-based registration methods are basically divided into three categories: model parameter methods [10], corner matching methods [11], [12], [13], and unsupervised methods [14]. The model parameter methods utilize a trained deep neural networks predicts matching labels of patch-pairs from the sensed and reference images. Afterward, a homography matrix for image registration is computed according to the matched points. Although the output of the method is more direct, it is time consuming to train the network with new images. Corner matching method is consistent with the traditional feature-based matching method. The CNNs perform three steps together: the features detection, orientation calculation, and representation extraction. Compared with the model parameter methods, the corner matching methods reduce the difficulty of network model training, and take the advantage of neural network in feature extraction to increase the accuracy of corner matching, so as to improve the registration accuracy. Unsupervised learning strategies rely on similarities to judge the registration. However, because of complex terrain environments in high-resolution remote sensing images, the existing standards of similarity measures (e.g., cross-correlation, mutual information index) are not a good measure for the similarity of two images, which directly affect the performance of unsupervised models [14].

Owing to complicated scenes with notable features changes of ground objects, and lack of ground truth registration datasets for high-resolution remote sensing images, the registration performance of the above learning approaches will substantially lessen. Accordingly, we propose a novel R-net based on corner matching strategies to tackle those troubles. R-net is a CNN with a U-shape structure, which adopts a corner loss function for extracting robust invariant features, and outputs corner points and corresponding descriptors for improving registration reliability. As for the lack of registration datasets, we propose an automatic annotation strategy inspired by Bagging theory, which yields credible registration datasets, simulating projection changes caused by different camera shooting angles.

B. Change Detection

Methods of building change detection of remote sensing images can be divided into pixel-based [15], [16], [17], feature-based [18], [19], [20], object-oriented-based [21], [22], [23], and deep learning-based approaches [20], [23], [24]. The pixel-based algorithms only consider the change of the current pixel, which may slightly affect the accuracy of classification of low-resolution images. However, for high-resolution images, a single pixel cannot contain the whole ground object information, and the accuracy of classification will be greatly reduced [17]. Feature-based methods only obtain a small number of features, which are generally for specific application directions, and has insufficient algorithm robustness to disturbance of noise environment [20]. Object-oriented methods adopt the scheme of first classification and then detection, which effectively inhibits the influence caused by some pseudochanges of ground objects, such as uneven shadow and illumination. However, it is easy to be disturbed by camera shooting angle or indistinguishable moving target in high resolution remote sensing image [23]. The pixel-based, feature-based, and object-oriented-based approaches belong to traditional algorithms with low complexity and fast running speed, and have good performance in low-resolution images, but are not robust enough to tackle diverse complex scenes.

Recently, owing to the robust feature extraction and discriminative capabilities of CNNs and transformer-based models, they have been widely applied in processing remote sensing images. In the task of building change detection in remote sensing images, network models based on CNN and transformer architectures have achieved state-of-the-art performance. Chen and Shi [25] proposed a change detection model named STANet, which integrates a basic spatial-temporal attention module (BAM) and pyramid spatial-temporal AM (PAM), BAM uses global spatiotemporal relationships to obtain better discriminant features, while PAM has multiscale attention representation to obtain finer details. The transformer, initially proposed in 2017 and primarily used in natural language processing. Since the introduction of the vision transformer by Dosovitskiy et al. [26], transformer models have been extensively utilized in computer vision (CV). Due to the transformer's strong representational power, it has demonstrated performance comparable to or better than CNN-based models across various CV tasks, such as classification [27], [28], object detection [29], [30], segmentation [31], and change detection [32], [33]. Similarly, transformer models have been extensively used in change detection tasks in remote sensing images. Chen et al. [32] introduced the bitemporal image transformer (BIT) network model, which combines transformer and CNN networks for change detection tasks. Utilizing a CNN backbone, advanced semantic features are extracted from input

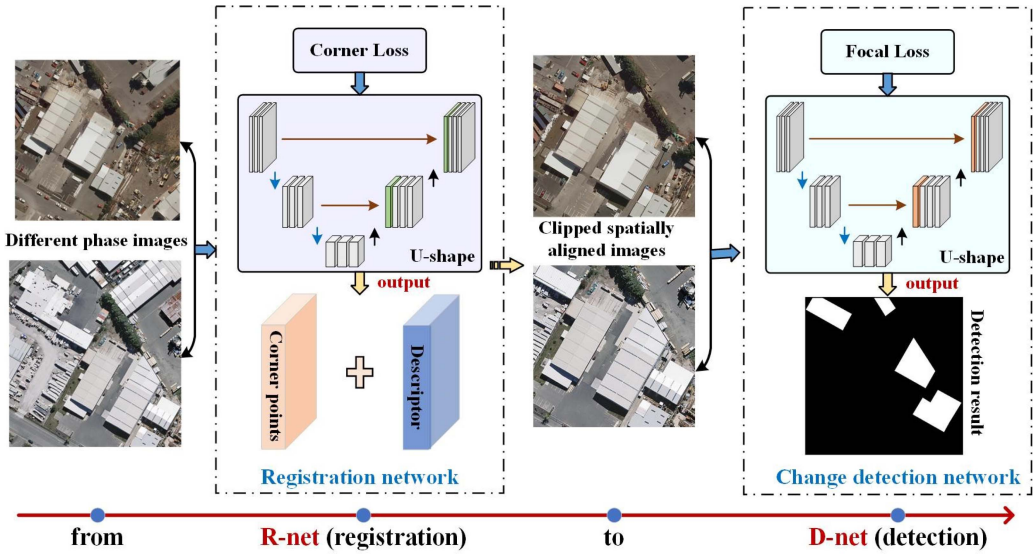

Fig. 1. Framework of the R&D-net for the building change detection in high-resolution remote sensing images acquired in different temporal phases.

images. Spatial attention mechanisms map each temporal feature transformation to a semantic token set, followed by transformer modeling of remote context in bitemporal images, further enhancing the model's feature-capturing ability. Meanwhile, Bandare and Patel [33] argued that CNN dependency is unnecessary. They propose using a lightweight transformer decoder within siamese networks to effectively perform change detection tasks. These models have excellent performance in the task of building change detection, but neglect the evaluation of building profile extraction performance.

By comparison, the deep learning-based approaches can learn effective and robust features in remote sensing images for avoiding some drawbacks met in traditional approaches, which is exactly suitable for change detection. Nevertheless, complicate scenes and interferences potentially arouse missing or wrong segmentations, especially for high-resolution remote sensing images. We propose a novel change D-net to solve those difficulties. D-net integrates the pyramid pooling operation and a focal loss function for accommodating various complex scenes, different object sizes, and image resolutions.

III. Method

A. Framework of R&D-Net

The framework of R&D-net for the building change detection in high-resolution remote sensing images is shown in Fig. 1, which contains a R-net and a change detection network D-net. The R-net takes two images acquired in different temporal phases as inputs and then outputs corner points and corresponding descriptors. Through a transformation, spatially aligned images are obtained. After that, the images are fed into the D-net for detecting building changes.

B. Registration Network

The R-net is on the basis of corner matching strategies, aiming to search matching corners in different phase images, where the matching corners have the same geographical coordinates. The structure of R-net is shown in Fig. 2. The coding part of R-net consists of convolution layer, feature normalization layer, activation function layer, and pooling layer. Three maximum pooled layers are passed during the activation function using leaky ReLU. The decoding part outputs the feature image of the same size as the input image through three up-sampling. The input of each decoding contains two parts: one is the feature map corresponding to the encoding structure, and the other is obtained from upsampling the feature map of the previous layer. Accordingly, the output of R-net is a probability graph with size of $H \times W \times 2$, representing the probabilities of whether it is a corner. Through the encoding and decoding network, we get a feature tensor. When the tensor flows through the corner detection branch, it will first pass through a convolution layer with the size of the convolution kernel of 1×1, then flow forward and pass through the softmax layer. Nonmaximum suppression (NMS) is applied to delete the surrounding redundant corner points.

There is a description vector branch in R-net. When two corners have the same geographical coordinates, the norm distance between the corresponding description vectors should be as small as possible. Otherwise, it should be as large as possible. The description vector branch is a feature tensor with size of $H \times W \times 256$. The vector that flows into the branch goes through a convolution layer with 1×1 convolution kernel. After that, the feature tensor will pass through an L2 norm structure for normalizing the feature vector.

The loss function in R-net consists of two parts: one is the loss of corner detection L_p, and the other is the loss of description vector L_d. The parameter λ is used to balance the weight between them, so that both losses can be optimized at the same time during training. The definition of the total loss function can be described as

$$L\left(P, P', D, D'; \hat{P}, \hat{P}', H\right) = L_p\left(P; \hat{P}\right) + L_p\left(P'; \hat{P}'\right) + \lambda L_d\left(D, D'; H\right) \quad (1)$$

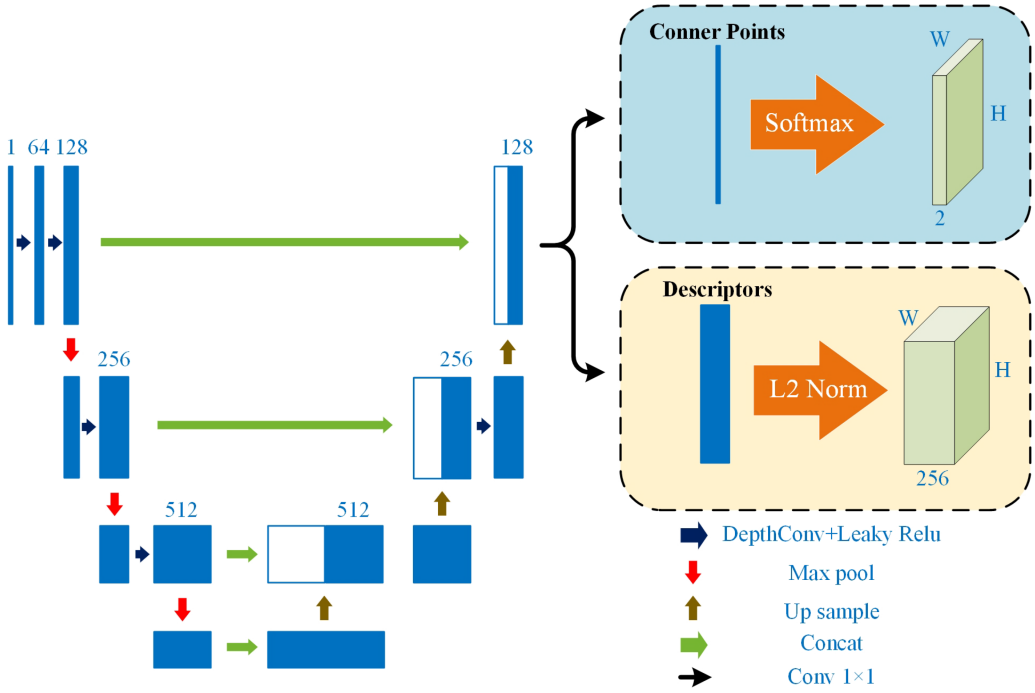

Fig. 2. Diagram of R-net. R-net adds a description vector branch on a CornerNet network, which outputs the description vector corresponding to the corner.

where L is the total loss function, P and P' are the corner probability graphs of the two images, D and D' are the description vector feature graphs of the two images, P' and $\hat{P}'$ is the true labels of corner points in the two images, H is the homography matrix, and λ is a constant, respectively.

The loss of corner detection L_p is defined as

$$L_p\left(P;\hat{P}\right)=\frac{1}{H\times W}\sum_{x=0,y=0}^{H-1,W-1}f_F\left(P_{x,y};\hat{P}_{x,y}\right)$$

$$f_F\left(P_{x,y};\hat{P}_{x,y}\right)=-\sum_{c=0}^{c-1}\alpha\hat{P}_{x,y,c}(1-P_{x,y,c})^{\gamma}\log P_{x,y,c} \quad (2)$$

where H and W denote the height and width of the input image, x and y denote the coordinates of the corner probability graph, c denotes the channel number of the probability graph, γ denotes the modulation coefficient, and α denotes the weight parameter of positive and negative samples.

The loss of description vector L_d is defined as

$$\begin{aligned}&L_p\left(D,D';H\right)\\&=\frac{1}{(H\times W)^2}\sum_{x=0,y=0}^{H-1,W-1}\sum_{x'=0,y'=0}^{H-1,W-1}f_h\left(D_{x,y},D'_{x',y'};H\right)\\&f_h\left(D_{x,y},D'_{x',y'};H\right)\\&=s\times\max\left(0,m_p-D_{x,y}^{T}D'_{x',y'}\right)\\&\quad+\lambda_d\times(1-s)\times\max\left(0,D_{x,y}^{T}D'_{x',y'}-m_n\right)\end{aligned}$$

$$s=\begin{cases}1,\text{if flag}\le 4\\0,\text{Other}\end{cases}$$

$$\text{flag}=\left\|\text{Coord}(x,y)-F_H\left(\text{Coord}\left(x',y'\right),H\right)\right\|_2 \quad (3)$$

where D and D' denote the description feature graphs for the first and second images. f_h is the hinge loss function, whose truncation characteristic makes all simple samples do not participate in the final hyperplane decision, greatly reducing the dependence on the number of training samples and improves the training efficiency. λ_d is a parameter, which balance the positive and negative examples. m_p and m_n are positive and negative soft intervals, which is beneficial to improve the robustness of the model. s is the positive and negative sample marker variable, which is determined by whether the corner points on the two images have the same geographic coordinates. $\text{Coord}(\cdot)$ represents the coordinate function, where $\text{Coord}(x,y)$ and $\text{Coord}(x',y')$ are the point coordinates in the first image and the second image, respectively. $f_H(\cdot)$ is the homography transformation function and H is the homography matrix of the first image. The point on the second image is converted to the point on the first image through the f_H function, and the two points have the same geographic coordinates. $\|\cdot\|_2$ represents a two-norm. When the distance between the corner point $f_H(\text{Coord}(x',y'),H)$ on the second image and the corner point $\text{Coord}(x,y)$ on the first image is less than or equal to 4, the two points are considered to be matching. In this case, the mark s is 1, that is, the positive sample. Otherwise, the mark s is 0, that is the negative sample.

C. D-Net

A novel change detection network (viz., D-net) is introduced in this section, whose network structure is shown in Fig. 3. The D-net adopts classic encoding and decoding structures. In the encoding part, ResNet50 [34] is used as the backbone network.

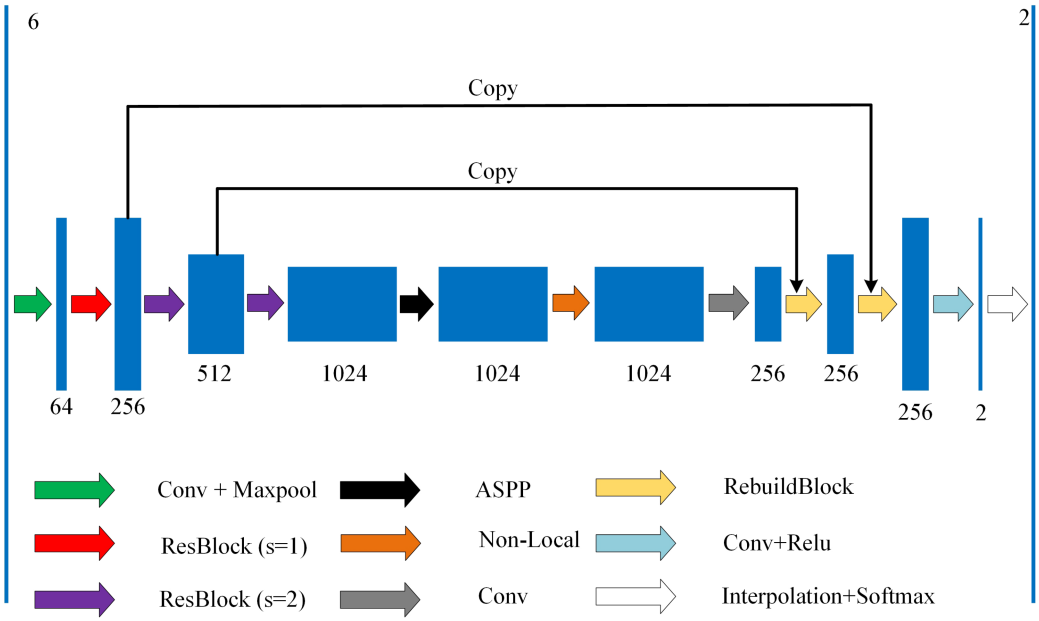

Fig. 3. Diagram of D-net.

First, we use a convolution layer with a step size of 2 and a maximum pooling layer with a step size of 2 to eliminate the redundant information of feature maps, reduce the number of features need to be processed, and speed up the forward reasoning. The convolution kernel size is 7×7, which can expand the receptive field of the image and help to fully collect the image information. Second, three residual structures (i.e., ResBlock) are adopted, whose number of convolution layers is 9, 12, and 18, respectively. Large number of convolution layers cannot only provide powerful ability of fitting, but also fully mine the deep semantic information in the image. The ResBlock module is based on the ResNet network, which consists of several block-based modules. The bulk part of the Bottleneck module includes three convolutional layers with sizes of 1×1, 3×3, and 1×1. Accordingly, the size of feature maps output by the encoding part is $\frac{H}{16} \times \frac{W}{16} \times 1024$, where 1024 is the number of feature channels, and H and W denote the height and width of the input image. The atrous spatial pyramid pooling (ASPP) structure is used to obtain the feature information of different scales. A nonlocal attention model is adopted between the encoding and decoding parts. The input and output feature maps have the same size. In the decoding part, we introduce a RebuildBlock module that effectively use the feature maps to retrieve the position information and reconstruct feature maps with edge information. The structure of RebuildBlock module is shown in Fig. 4, where the Skip Map represents the feature map copied from the encoding structure and the Map is the feature map propagated forward. The Skip Map first flows through a convolutional layer and then flows into a bottleneck-base residual model. The Map will make the size of its feature map consistent with that of Skip Map through the upsampling. Then, we add the two branches directly to the counterpoint, fuse the feature maps through a bottleneck-base residual model, and output results. The RebuildBlock module is used twice in the network to perform two up-samples on the feature map, and then the width and height of the feature map are $\frac{H}{4} \times \frac{W}{4}$ (the channel number is compressed to two through a convolution layer), as shown in Fig. 3.

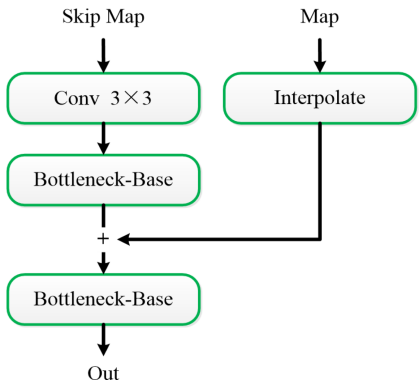

Fig. 4. Structure of RebuildBlock module.

In the training process, the number of positive samples is far less than that of negative samples. In this case, the model will pay more attention to negative samples during optimization, which leads to low sensitivity of the trained model to the changed area during testing, and tends to predict as unchanged. Therefore, we use focus loss function to optimize this problem, which can be defined as

$$\mathrm{FL}(y;\hat{y}) = -\alpha\hat{y}(1-\hat{y})^{\gamma}\log y \tag{4}$$

where α denotes a weight parameter of positive and negative samples.

IV. Experimental Results

In this section, we present experimental results on change detection datasets. The performance of our method is verified

TABLE I
THREE PUBLIC DATASETS OF HIGH-RESOLUTION REMOTE SENSING IMAGES

DataSet	LEVIR_CDs	WHU_BCD	CD_Data_GZ
Spatial resolution (meter/pixel)	0.5	0.3	0.55
Size (pixels)	1024 × 1024	32 507 × 15 354	1006 × 1168 ∼ 4936 × 5224
Image size	512 × 512	512 × 512	512 × 512
Number of pairs	2548	1827	729

by comparison experiment and ablation experiment, for a comprehensive analysis of the proposed method.

A. Datasets, Baseline Methods, and Metrics

Datasets: The dataset should contain images before and after the change. The qualified images in LEVIR_CDs and WHU_BCD are taken as the training set and verification set of this experiment, and the images in CD_Data_GZ are taken as the test set. All images are cut into patches with size of 512 × 512. As given in Table I, 2548 pairs of images are obtained from LEVIR_CDs, 1827 pairs of images are obtained from WHU_BCD, and 729 pairs of images are obtained from CD_Data_GZ. The number of image pairs in training set, verification set and test set is 4000, 375, and 729.

Baseline methods: For R-net, we adopt four corner detection methods as baseline methods for comparisons of registration performance (i.e., SIFT [35], ORB [36], Lift [13], and SuperPoint [14] methods). Among them, SIFT and ORB are traditional methods, which are implemented with OpenCV. Lift and SuperPoint are deep learning-based methods, and we used the training model provided by the authors. To fully evaluate D-net's performance, We selected SegNet [37], U-Net [38], FC-EF [39], FC-Siam-Di [39], FC-Siam-Conc [39], DeeplabV3 [40], DeeplabV3+[41], STANet [25], BIT [32], and ChangeFormer [33] models as comparison models. These models have excellent performance in building change detection tasks. The backbone network of SegNet, U-Net, FC-EF, FC-Siam-Di, and FC-Siam-Conc models is VGG, DeeplabV3, DeeplabV3+, and STANet, and the backbone network of our model is ResNet. The backbone of BIT and ChangeFormer is transformer.

Metrics: For R-net, we adopt four metrics for quantitative comparisons, that is, corner repeatability (Rep), average accuracy of corner matching (Match mAP), match the positioning error (Match mLE), and validity of homography estimation (Hom). Assume that N_1 corner points are detected on the first image and N_2 corner points are detected on the second image, the formula for defining Rep is defined as

$$\text{Rep} = \frac{1}{N_1+N_2}\left(\sum_{i=1}^{N_1}\text{Corner}R(X_i)+\sum_{j=1}^{N_2}\text{Corner}R\left(X_j'\right)\right)$$

$$\text{Corner}R(X_i) = \left(\min_{j\in 1_1,\ldots,N_1}\|f_H(X_i)-X_j'\|_2\right)\le\varepsilon \quad (5)$$

where Corner$R(X_i)$ indicates whether corner point X_i in the first corner set can be found in the second corner set with a distance less than or equal to ε. If it can be found, it will return 1; otherwise, it will return 0. $f_H(\cdot)$ is a homography transformation function. The range of Rep is [0,1], and the larger the value is, the better the algorithm performance is.

The definition of Match mAP and Match mLE can be described as

$$\text{Match mAP}(X_i,X_j) = \|f_H(X_i)-X_j\|_2$$

$$\text{Match mLE}(X) = \frac{1}{K}\sum_{i,j:\text{Match}(X_i,X_j)}\|X_i-X_j\|_2 \quad (6)$$

where (X_i, X_j) is a pair of matching corner points, $f_H(\cdot)$ is a real homography transformation matrix between two images, K is the number of correct matches, and Match(X_i, X_j) denotes the selection of corner pairs with correct matches for calculation, respectively. The range of Match mAP is [0,1], and the larger the value, the better the algorithm performance. If K is 0, the positioning error of corner matching is ε. Then, the range of corner matching positioning error is $[0, \varepsilon]$. The smaller the value is, the smaller the positioning error is.

First, randomly select four points on the first image and define them as $C_1, C_2, C_3,$ and C_4. Then, using the estimated matrix and the true matrix to transform these four points to the second image, resulting in C_1', C_2', C_3', and C_4', and $\hat{C}_1$, $\hat{C}_2$, $\hat{C}_3$, and $\hat{C}_4$. Calculate the difference between the two sets of points, and repeat this process N times. The Hom is then defined as

$$\text{Hom} = \frac{1}{N}\sum_{i}^{N}\left(\left(\frac{1}{4}\sum_{j}^{4}\|C_{ij}'-\hat{C}_{ij}\|_2\right)\le\varepsilon\right) \quad (7)$$

where C_{ij}' represents the jth point in the ith time and ε represents the threshold value. The range of Hom is [0,1], and the larger the value is, the better the algorithm performance is.

For D-net, the regional similarity, profile accuracy, parameter number, and frame per second (FPS) as four metrics. Regional similarity(J) can measure the coincidence degree between the segmented and real areas, which is defined as

$$J = \frac{M\cap G}{M\cup G}$$

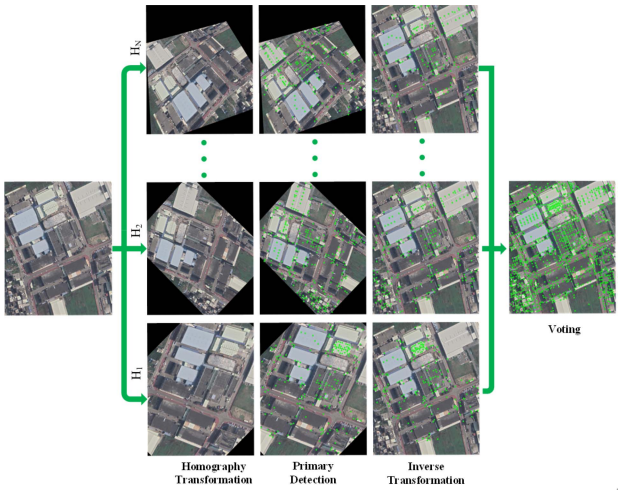

Fig. 5. Procedure of multiview labeling strategy, which consists of the homography transformation, primary detection, inverse transformation, and voting.

$$M \cap G = \sum_{x}^{W} \sum_{y}^{H} M_{x,y} \times G_{x,y}$$
$$M \cup G = \sum_{x}^{W} \sum_{y}^{H} (M_{x,y} + G_{x,y}) > 0 \tag{8}$$

where M represents the predicted mask matrix, and G represents the true label matrix. The element in the matrix is 0,1, where 0 denotes no change, and 1 denotes the change region. The height and width of M and G are H and W.

The profile accuracy metric (F) to evaluate the performance of building change detection is defined as

$$F = \frac{2P_c R_c}{P_c + R_c}, P_c = \frac{TP}{TP + FP}, R_c = \frac{TP}{TP + FN}$$
$$TP = \sum_{x}^{W} \sum_{y}^{H} M_{c_{x,y}} \times G_{c_{x,y}}$$
$$TP + FP = \sum_{x}^{W} \sum_{y}^{H} M_{c_{x,y}}$$
$$TP + FN = \sum_{x}^{W} \sum_{y}^{H} G_{c_{x,y}} \tag{9}$$

P_c represents the precision of the profile, R_c represents the recall of the profile, M_c represents the predicted mask matrix, and G_c represents the true label matrix. TP represents the number of accurately predicted profile pixels, $TP + FP$ represents the number of predicted profile pixels, and $TP + FN$ represents the number of true profile pixels.

B. Registration Comparisons

Inspired by Bagging theory, we propose an automatic multiview labeling method to obtain labeling corners, as shown in Fig. 5. First, N homography transformation matrices randomly generated are implemented to original images (the matrices indicate N different perspectives). The transformed results are shown in the second column. Second, the transformed results are input a primary model, and preliminary detected results are obtained accordingly, as shown in the third column. Third, the inverse homography transformation is applied to project the detected points back to the original coordinates. The fourth column denotes the projected results, that is, N weak classification results. Finally, a voting strategy is used to fine the results, whose votes are greater than a threshold, as shown in the rightmost column.

The following settings are given in registration comparisons, aiming to ensure the fairness.

1) For SIFT, ORB, LIFT, SuperPoint, and R-net, the maximum number of corner points should not exceed 1000 during detection.
2) The parameters of NMS are the same for the refinement of corner detection results.
3) In the evaluation of Hom, the corner filtering algorithms and parameters are the same all algorithms.
4) A threshold ε is used to determine whether the corners are correctly matched. The ε is set to 4 during the comparisons.
5) The environment is the same for each algorithm, that is, CPU: Intel Xeon Sulver 4110 CPU @2.10 GHz, RAM: 128 G DDR4 3.2 GHz, GPU: NVIDIA RTX 2080Ti, VRAM: 11 G, HDD: SSD 4 T, Python: version 3.7.7, PyTorch: version 1.7.1, and OpenCV: version: 4.5.1.

Table II lists the quantitative comparison results obtained using by the SIFT, ORB, LIFT, SuperPoint, and R-net. R-net is only slightly behind ORB on Rep and only behind SIFT on Match mLE. However, R-net is significantly ahead of other baselines in Match mAP and Hom. Specifically, R-net not only has much higher Match mAP index than that of the traditional methods (e.g., SIFT and ORB), but also improves by 5.9% compared with deep learning-based methods (e.g., LIFT SuperPoint). Similarly, R-net improved by 4.9% on Hom. This indicates that R-net has better corner matching accuracy and better registration performance than other baselines.

C. Change Detection Comparions

In the experiments of building change detection, the size of the image is 512 × 512, and the threshold value is set at 0.5 in prediction. ResNet50 pretrained on ImageNet was used as the backbone of the model, the batchsize of the model was set to 8, the initial learning rate of the model was set to 0.005, and the learning rate during the training process was adjusted by warmup training strategy. The training epoch set to 200. The other models keep their original Settings. The model training are performed on the same RTX2080ti graphics card in this study. As given in Table III, D-net model achieves suboptimal and optimal performance in regional similarity and profile accuracy, respectively. The regional similarity of D-net is 86.7, trailing only behind ChangeFormer, which achieved 87.4%. It attains a Profile accuracy of 81.9%, surpassing all other baseline. In addition, FC-EF, FC-Siam-Di, and FC-Siam-Conc have the fewest

TABLE II
QUANTITATIVE REGISTRATION COMPARISONS OF SIFT, ORB, LIFT, SUPERPOINT, AND R-NET ALGORITHMS

Method	Rep	Match mAP	Match mLE	Hom	FPS
SIFT[35]	49.6	71.4	**0.892**	71.3	13.1 (cpu)
ORB[36]	**64.7**	64.9	1.183	64.5	70.4 (cpu)
LIFT[13]	48.8	72.0	1.150	72.9	13.5 (cpu), 21.7 (gpu)
SuperPoint[14]	57.5	80.5	1.146	82.2	8.3 (cpu), 23.6 (gpu)
R-net	64.1	**86.4**	0.997	**87.1**	21.5 (cpu), 70.9 (gpu)

Bold entities represent the best results.

TABLE III
COMPARATIVE EXPERIMENT OF DIFFERENT CHANGE DETECTION MODELS

Methods	Backbone	Regional similarity	Profile accuracy	Para.	FPS
SegNet[37]	VGG16	66.8	53.1	112.4M	32.8
U-Net[38]	VGG13	75.1	74.8	65.9M	31.4
FC-EF[39]	VGG13	77.4	70.4	**1.35M**	**142.9**
FC-Siam-Di[39]	VGG13	80.7	74.2	**1.35M**	104.1
FC-Siam-Conc[39]	VGG13	77.9	71.5	**1.35M**	98.7
Deeplab V3[40]	Resnet50	80.5	71.2	149.1M	43.4
Deeplab V3+[41]	Resnet50	83.9	76.0	154.0M	24.1
STANet[25]	Resnet50	85.5	76.1	16.93M	43.5
BIT[32]	Transformer	86.2	79.8	-	-
ChangeFormer[33]	Transformer	**87.4**	79.8	-	-
D-net	Resnet50	86.7	**81.9**	113.5	53.6

Bold entities represent the best results.

TABLE IV
ABLATION EXPERIMENT

RebuildBlock	ASPP	Non-local	Regional similarity	Profile accuracy
✓	✓	✓	**86.7**	**81.9**
✓	✓	✗	85.6	81.0
✓	✗	✓	83.5	80.3
✓	✗	✗	82.2	77.5
✗	✓	✓	74.2	67.1
✗	✓	✗	72.9	65.8
✗	✗	✓	70.4	64.0
✗	✗	✗	68.7	62.8

Bold entities represent the best results.

parameters, with FC-EF exhibiting the best FPS performance. Compared with other backbones, D-net's performance in terms of parameters and FPS is middling, neither excelling nor lagging behind. This is acceptable considering the excellent performance of D-net. Overall, D-net has better building change detection performance.

D. Ablation Experiment

In order to further study the effect of RebuildBlock, ASPP and nonlocal modules on the model performance, some ablation experiments are carried out. The model achieves optimal performance when equipped with RebuildBlock, ASPP, and nonlocal modules, as demonstrated in Table IV. In this case, the regional similarity reaches 86.7% and profile accuracy reaches 81.9%. Without RebuildBlock, both regional similarity and profile accuracy experience a significant decrease of 12.5% and 14.8%, respectively. This substantial decline can be attributed to the crucial role played by RebuildBlock in the D-net decoder module, which effectively integrates features from both encoder and decoder components to capture high-level semantic information from deep layers, as well as edge and profile details from shallow layers. The absence of ASPP results in a reduction of 3.2% in regional similarity and 1.6% in profile accuracy for the model since ASPP module plays a vital role in integrating multiscale features and significantly enhancing overall performance. Similarly, without incorporating nonlocal module into the model architecture, there is a decrease of 1.1% in regional similarity and 0.9% in profile accuracy due to its ability to facilitate global feature learning process thereby improving overall model performance.

V. DISCUSSION

The comparison experiment with baseline shows that our model has excellent performance. As shown in Fig. 6, models, such as SegNet and U-Net, which are used for general segmentation tasks, are prone to missegmentation when processing remote sensing images with complex ground object backgrounds, and are easy to missegment roads, cars, and other objects similar to buildings into buildings. DeeplabV3 and DeeplabV3+ have better segmentation performance than SegNet and U-Net, but there are also cases of missing segmentation and false segmentation. As shown in g, h, and i in Fig. 6, DeeplabV3 and DeeplabV3+ both mistakenly divide the road in h into buildings. FC-EF, FC-Siam-Di, FC-Siam-Conc, STANet, BIT, and ChangeFormer are models for building change detection, and our model has comparable performance with these models. FC-EF, FC-Siam-Di, and FC-Siam-Conc have similar change detection performance, and STANet have better profile detection performance than the models of the FC series, as shown in b and d

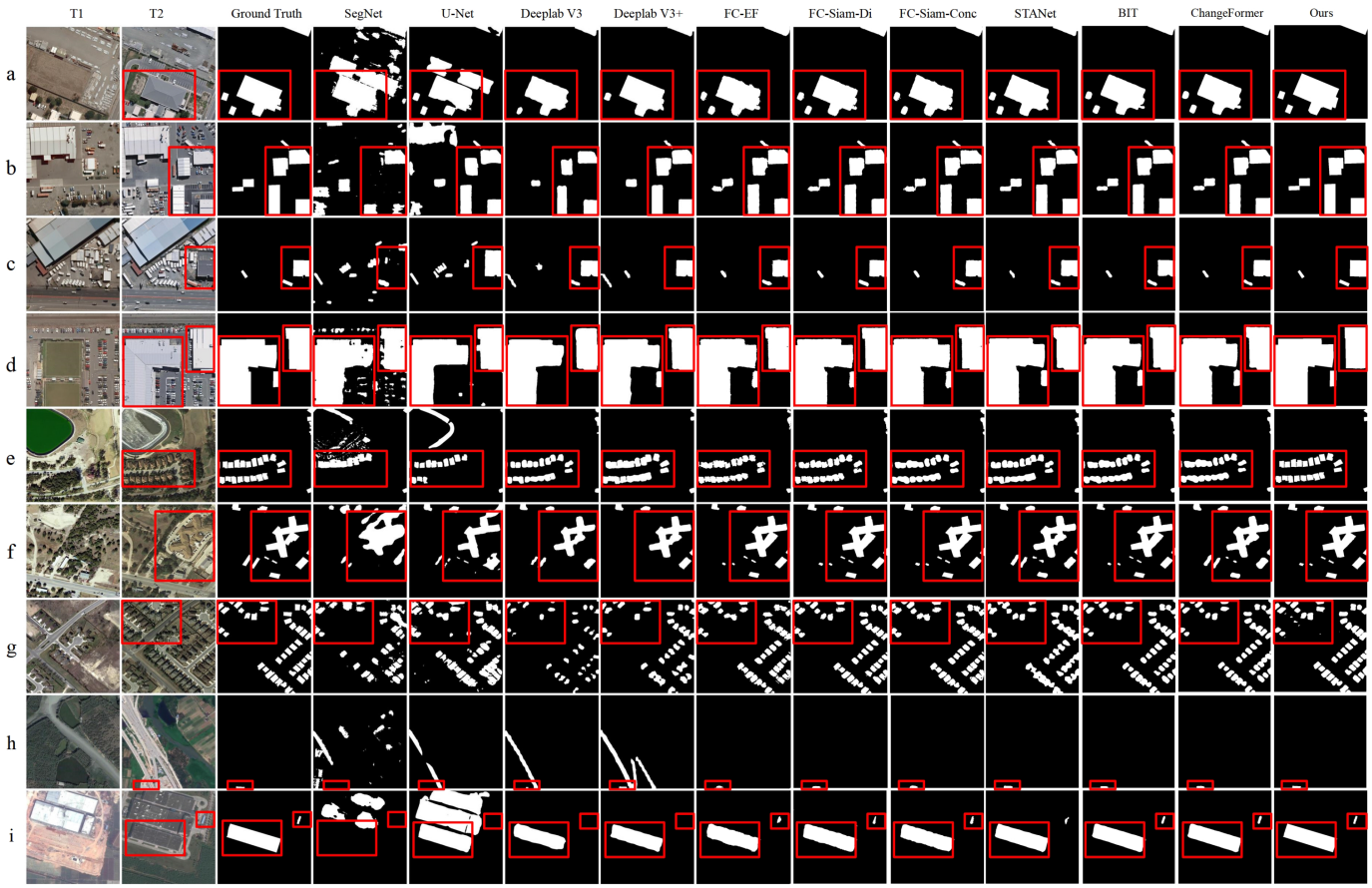

Fig. 6. Original remote sensing images, ground truths of building change, and detection results by using SegNet, U-Net, Deeplab V3, Deeplab V3+, FC-EF, FC-Siam-Di, FC-siam-Conc, STANet, BIT, ChangerFormer, and D-net.

in Fig. 6. Thanks to transformer's ability to learn global features of images, BIT and ChangeFormer have excellent performance and better detection performance of building changes, as shown in a and d in Fig. 6. In particular, our model has a better segmentation performance on the profile of the building and is closer to the ground truth, as shown in e, f, and i in Fig. 6. This is due to the presence of the RebuildBlock module in the model, which combines the features of the encoder and the decoder to enhance the model's learning of shallow features, such as building edges and profile. However, our model also has a small number of missing segmentation and false segmentation problems, such as missing segmentation in b and false segmentation in g. Future work will focus on improving the model's ability to learn global features to effectively solve this problem. Overall, our model has equally good change detection performance and better building edge detection performance compared with these baselines.

VI. Conclusion

In this article, we proposed R&D-net to detect building changes of high-resolution remote sensing images, aiming to deal with difficulties caused by complex ground object environment, vehicles, temporarily stacked debris, vegetation, roads, and other interference targets. The R&D-net including a R-net followed by a D-net. A method for multiview automatic labeling is proposed to address the issue of lacking a remote sensing image registration dataset. This method enables corner points labeling of remote sensing images without the need for manual supervision, significantly improving the accuracy of corner point labeling. To enhance the robustness of image pair registration, this article introduces the R-net structure. Compared with the baseline, the remote sensing image registration method based on the R-net exhibits better registration accuracy. tis article proposes the D-net for achieving change detection of buildings. To address the issue of incomplete segmentation in large-scale building change areas and missed segmentation of small-scale buildings, the ASPP module structure is effectively employed to alleviate this problem. To mitigate the problem of misjudgment in change detection, this article utilizes the nonlocal module to enable the model to capture global image information, thereby avoiding cases of misjudgment. Compared with baseline, the model has the second best regional similarity and the best profile accuracy. Experimental results demonstrate that the R&D-net cannot only extract robust invariant features for improving registration accuracy, but also improve both leaky and wrong segmentations under complex scenes. In general, the proposed R&D-net has a good comprehensive performance in the building change detection.

REFERENCES

[1] Y. Tang, X. Huang, and L. Zhang, "Fault-tolerant building change detection from urban high-resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1060–1064, Sep. 2013.

[2] Z. Yan, R. Huazhong, and C. Desheng, "The research of building earthquake damage object-oriented change detection based on ensemble classifier with remote sensing image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 4950–4953.

[3] L. Matikainen, J. Hyyppä, E. Ahokas, L. Markelin, and H. Kaartinen, "Automatic detection of buildings and changes in buildings for updating of maps," *Remote Sens.*, vol. 2, no. 5, pp. 1217–1248, 2010.

[4] H.-M. Chen, P. K. Varshney, and M. K. Arora, "Performance of mutual information similarity measure for registration of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11, pp. 2445–2454, Nov. 2003.

[5] Y. Bentoutou, N. Taleb, K. Kpalma, and J. Ronsin, "An automatic image registration for applications in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 9, pp. 2127–2137, Sep. 2005.

[6] S. Chen, X. Li, L. Zhao, and H. Yang, "Medium-low resolution multisource remote sensing image registration based on sift and robust regional mutual information," *Int. J. Remote Sens.*, vol. 39, no. 10, pp. 3215–3242, 2018.

[7] L. Lucchese, S. Leorin, and G. M. Cortelazzo, "Estimation of two-dimensional affine transformations through polar curve matching and its application to image mosaicking and remote-sensing data registration," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3008–3019, Oct. 2006.

[8] C. Lyu and J. Jiang, "Remote sensing image registration with line segments and their intersections," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 439.

[9] W. Ma, Y. Wu, Y. Zheng, Z. Wen, and L. Liu, "Remote sensing image registration based on multifeature and region division," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1680–1684, Oct. 2017.

[10] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, and L. Jiao, "A deep learning framework for remote sensing image registration," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 148–164, 2018.

[11] R. Fan, B. Hou, J. Liu, J. Yang, and Z. Hong, "Registration of multiresolution remote sensing images based on l2-Siamese model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 237–248, Dec. 2020.

[12] Y. Liu, X. Gong, J. Chen, S. Chen, and Y. Yang, "Rotation-invariant Siamese network for low-altitude remote-sensing image registration," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5746–5758, Sep. 2020.

[13] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 467–483.

[14] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.

[15] A. Varshney, "Improved NDBI differencing algorithm for built-up regions change detection from remote-sensing data: An automated approach," *Remote Sens. Lett.*, vol. 4, no. 5, pp. 504–512, 2013.

[16] C. Beumier and M. Idrissa, "Building change detection from uniform regions," in *Proc. Prog. Pattern Recognit. Image Anal. Comput. Vis. Appl. 17th Iberoamerican Congr.*, 2012, pp. 648–655.

[17] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1688.

[18] M. Janalipour and M. Taleai, "Building change detection after earthquake using multi-criteria decision analysis based on extracted information from high spatial resolution satellite images," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 82–99, 2017.

[19] N. Sofina and M. Ehlers, "Building change detection using high resolution remotely sensed data and GIS," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3430–3438, Aug. 2016.

[20] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1343.

[21] D. Argialas, S. Michailidou, and A. Tzotsos, "Change detection of buildings in suburban areas from high resolution satellite data developed through object based image analysis," *Surv. Rev.*, vol. 45, no. 333, pp. 441–450, 2013.

[22] S. Tanathong, K. T. Rudahl, and S. E. Goldin, "Object oriented change detection of buildings after the Indian Ocean Tsunami disaster," in *Proc. 5th Int. Conf. Elect. Eng./Electron. Comput. Telecommun. Inf. Technol.*, 2008, pp. 65–68.

[23] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-siamnet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 484.

[24] W. Liu et al., "Building footprint extraction from unmanned aerial vehicle images via PRU-Net: Application to change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2236–2248, 2021.

[25] 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.

[26] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, Oct. 2020.

[27] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12009–12019.

[28] Z. Tu et al., "Maxvit: Multi-axis vision transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 459–479.

[29] Z. Zong, G. Song, and Y. Liu, "Detrs with collaborative hybrid assignments training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 6748–6758.

[30] Z. Meng, X. Xia, R. Xu, W. Liu, and J. Ma, "HYDRO-3D: Hybrid object detection and tracking for cooperative perception using 3D LiDAR," *IEEE Trans. Intell. Veh.*, vol. 8, no. 8, pp. 4069–4080, Aug. 2023.

[31] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–120 90.

[32] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5607514.

[33] 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.

[34] 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.

[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[36] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[37] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Intervn. 18th Int. Conf.*, 2015, pp. 234–241.

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

[40] J. Chen, H. Feng, K. Pan, Z. Xu, and Q. Li, "An optimization method for registration and mosaicking of remote sensing images," *Optik*, vol. 125, no. 2, pp. 697–703, 2014.

[41] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

Zhong Chen received the M.S. degree in physical electronics from the Science and Technology University of Huazhong, Wuhan, China in 2003, and the Ph.D. degree in cartography and geographic information systems from the Institute of Remote Sensing Application, Chinese Academy of Sciences, Beijing, China, in 2006.

From 2006 to 2008, he was a Postdoctoral Fellow with the Huazhong University of Science and Technology, Wuhan, where he is currently an Associate Professor. His research interests include image analysis, remote sensing image processing, and target recognition.

Tong Zheng received the B.S. degree in measurement and control technology and instruments and the M.S. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2020 and 2023, respectively.

He is currently with Huawei Technologies, Company, Ltd., Chengdu, China. His research interests include deep learning and remote sensing image processing.

Junsong Leng received the B.S. degree in automation and the M.S. degree in mechanics from Huazhong Agricultural University, Wuhan, China, in 2019 and 2022, respectively. He is currently working toward the Ph.D. degree in control science and engineering with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan.

In 2022, he joined the State Key Laboratory of Multispectral Information Processing Technology, Huazhong University of Science and Technology. His research interests include computer vision, domain adaptation, and remote sensing image processing.

Jiahao Zhang received the B.S. degree in control science and engineering from the Wuhan University of Technology, Wuhan, China, in 2018, and the M.S. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, in 2021.

He is currently with Hikvision, Hangzhou, China. His research interests include computer vision and deep learning.

He Deng received the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2011.

He is currently a Professor with the Wuhan University of Science and Technology, Wuhan. His research interests include artificial intelligence and medical image processing.

Xiaofei Mi received the Ph.D. degree in signal and information processing from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2020.

She is currently an Assistant Professor with the National Engineering Laboratory of remote sensing satellite application, Beijing. Her research interests include land cover classification and change detection, including image processing and remote sensing application.

Jian Yang received the Ph.D. degree in cartography and geography information system from the Institute of Remote Sensing Application, Chinese Academy of Sciences, Beijing, China, in 2008.

He conducts research projects on the developing of the remote sensing Big Data high performance platform. His research interests include remotely sensing imagery processing and analyzing algorithms on high spatial resolution RS image, including multifeature extraction, multisource data fusion, land cover classification, and change detection of urban region.