AERNet: An Attention-Guided Edge Refinement Network and a Dataset for Remote Sensing Building Change Detection

Jindou Zhang, Zhenfeng Shao, Qing Ding, Xiao Huang, Yu Wang, Xuechao Zhou, and Deren Li

***Abstract*— Advancements in Earth observation technology enable the detection of surface changes in intricate urban environments. Building change detection (BCD) plays a crucial role in urban planning and environmental monitoring. However, existing deep learning-based BCD algorithms exhibit limited capability in feature extraction, feature relationship comprehension, sample imbalance mitigation, and accurate boundary identification for changed objects. To address these challenges, we introduce an attention-guided edge refinement network (AERNet) that uses a global context feature aggregation module (GCFAM) to aggregate information from extracted multilayer context features. Our approach incorporates an attention decoding block (ADB) guided by enhanced coordinate attention (ECA) to capture channel and location associations between features. Furthermore, we use an edge refinement module (ERM) to enhance the network's capacity to sense and refine the edges of changed areas. To tackle the issue of class imbalance and augment the algorithm's feature learning ability, we devise a novel self-adaptive weighted binary cross-entropy (SWBCE) loss function, combined with a deep supervision (DS) strategy. Experiments are conducted on two publicly available datasets, GDSCD and LEVIR-CD, and our newly developed high-resolution complex urban scene BCD dataset, i.e., HRCUS-CD. The latter dataset comprises 11 388 pairs of images at 0.5-m resolution and more than 12 000 labeled change buildings. Comparative experiments indicate that AERNet surpasses advanced competitive methods, while ablation experiments demonstrate the effectiveness of AERNet's model components and the SWBCE loss function. Efficiency comparison confirms that AERNet achieves comprehensive detection performance with superior effectiveness and robustness.**

***Index Terms*— Attention-guided edge refinement network (AERNet), building change detection (BCD), coordinate attention (CA), dataset, deep supervision (DS), edge refinement.**

I. Introduction

THE global population has continued to grow and urbanization has accelerated, which has continued to intensify the depletion of natural resources, such as land and minerals, and has placed higher demands on the sustainable development of societies. Iterative advances in remote sensing technology have made it more closely integrated with national economies and people's lives, and its role in urban expansion [1] and sustainable social development [2] has become increasingly important. Remote sensing change detection (CD) aims to identify change differences between dual-temporal or multitemporal remote sensing images (RSIs) in the same region [3], which is an important research direction in remote sensing technology and has been used extensively in urban planning [4], resource monitoring [5], and disaster emergency response [6]. With the continuation of urbanization, building CD (BCD) has drawn more and more interest, specifically in illegal building identification [7] and urban disaster assessment [8]. Since the 21st century, benefiting from the innovative development of various high-tech sensor devices, remote sensing observation technologies of the Earth have made significant progress with high-resolution RSIs becoming more easily available [9]. Although the detailed feature representation of objects from high-resolution images greatly facilitates the detection of building changes, BCD remains underexploited and deserves more attention.

The traditional BCD methods can be split into two groups: pixel-based approaches and object-based approaches. The pixel-based approaches produce the difference image by directly comparing information of image pixels and then dividing the change by threshold segmentation [10], and they include the change vector analysis method [11], spectral angle mapping method [12], principal component analysis [13], etc. Although pixel-based approaches are convenient to use, they neglect the spatial context information, which leads to the inability to obtain higher level object representation. The object-based approaches aim to segment RSIs into separate objects and apply the captured information to analyze the differences between images [14]. Example approaches include Markov random field method [15], sliding window statistical method [16], etc. They are able to better leverage rich

Manuscript received 3 April 2023; revised 11 May 2023 and 21 June 2023; accepted 23 July 2023. Date of publication 4 August 2023; date of current version 15 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42090012; in part by the Guangxi Science and Technology Plan Project under Grant Guike 2021AB30019; in part by the Hubei Province Key Research and Development Project under Grant 2022BAA048; in part by the Sichuan Province Key Research and Development Project under Grant 2022YFN0031, Grant 2023YFN0022, and Grant 2023YFS0381; in part by the Zhuhai Industry-University-Research Cooperation Project under Grant ZH22017001210098PWC; in part by the Shanxi Provincial Science and Technology Major Special Project under Grant 202201150401020; and in part by the Guangxi Key Laboratory of Spatial Information and Surveying and Mapping Fund Project under Grant 21-238-21-01. *(Corresponding author: Zhenfeng Shao.)*

Jindou Zhang, Zhenfeng Shao, Qing Ding, Yu Wang, Xuechao Zhou, and Deren Li are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zhangjindou@whu.edu.cn; shaozhenfeng@whu.edu.cn; dingqing@whu.edu.cn; wy2022@whu.edu.cn; 2021206190064@whu.edu.cn; drli@whu.edu.cn).

Xiao Huang is with the Department of Geosciences, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: xh010@uark.edu).

Digital Object Identifier 10.1109/TGRS.2023.3300533

image information, i.e., spectral, texture, and geometry, and effectively suppress the effect of pretzel noise. However, high-resolution RSIs' complex spectral and texture features and uncertain object segmentation create new difficulties [17]. Despite their many limitations, traditional methods still stand strong in specific remote sensing CD because they require fewer training samples.

Deep learning has demonstrated considerable benefit in RSIs applications over the past decade due to strong feature fitting capability, which include CD [18], image retrieval [19], and semantic segmentation [20]. Deep learning BCD methods can be broadly classified into two types: metric-based approaches and classification-based approaches.

The metric-based approaches determine whether a change has occurred by comparing the parametric distances of dual-temporal image pixel pairs, and L1 and L2 distance are frequently used to determine whether changes have taken place. For instance, STANet [21] used a pyramid attention mechanism to mitigate false detections due to alignment errors in dual-temporal images. DASNet [22] enhanced network's ability using a unique dual-attention mechanism and a newly designed loss function. DSAMNet [23] was a deeply supervised CD network, which focuses on the pseudochange and noise problems in the CD process. Scholars have tried different loss functions to obtain better accuracy, including contrast loss [24], triplet loss [25], etc. Parameter setting of the loss function requires extensive experimental validation, and there is still room for improving model's generalization ability on different datasets.

The classification-based approaches compare the results extracted from two images to obtain the change probability at the pixel level, which indicates the change in the feature at that pixel if a certain threshold is reached. For instance, Peng et al. [26] developed the algorithm using improved UNet++ to address the problem of error accumulation. Feature difference convolutional neural network (FDCNN) [27] applied the idea of feature difference and the newly proposed loss function to achieve good results. DSANet [28] used special feature extraction module and efficient network construction to effectively extract multilevel features. To reduce the dependence on the samples, single-temporal supervised learning (STAR) [29] CD network used only unpaired images to train a high-precision change detector. Bandara and Patel [30] designed a transformer-based Siamese network by transferring the relevant aspects of converters from the natural language processing (NLP) domain into the RSIs domain. Considering the building extraction results can be helpful for BCD, Liu et al. [31] applied multitask learning strategy to increase the detection area's integrity. Gao et al. [32] used the results of building extraction as a priori information and introduced the idea of refined boundary extraction. Despite the above efforts, the following challenges still exist in the BCD task.

1) The BCD models are insufficient in obtaining global context relationships between features and distinguishing between changed features in the decoding stage and suffer from sample class imbalance. The detection results tend to present notable false or missed detections, as well as irregularity building boundaries.
2) Large-scale BCD datasets are lacking in more difficult and complex urban environments. For training and testing, deep learning BCD algorithms need a lot of label data, as well as more benchmark data for performance judging.

To address these current problems, we propose an attention-guided edge refinement network (AERNet) for the BCD task. In addition, we develop a new, large-scale, open-sourced, high-resolution complex urban scene BCD (HRCUS-CD) dataset. The main contributions of this article are summarized as follows.

1) We introduce a novel BCD network, AERNet, which leverages attention mechanism and incorporates an edge refinement module (ERM) along with the deep supervision (DS) strategy to effectively detect changing buildings in RSIs. The experimental results from three datasets demonstrate AERNet's superior efficiency, detection performance, generalizability, and robustness.
2) We develop a self-adaptive weighted loss function that enhances the network's robustness by incorporating self-adaptive accuracy evaluation metrics as weight coefficients. This approach effectively mitigates the influence of sample class imbalances on the model's learning ability and improves the CD performance of the network. In addition, it exhibits commendable generalizability.
3) We create a high-resolution complex urban scene BCD dataset, named HRCUS-CD, which comprises 11 388 pairs of RSIs (256 × 256) at a 0.5-m resolution and more than 12 000 labeled change instances. This dataset expands the coverage and area of interest within the BCD domain, making it more comprehensive and valuable for practical applications.

The proposed AERNet code and HRCUS-CD dataset will be released at: https://github.com/zjd1836/AERNet. The rest of this article is organized as follows. Section II describes the proposed HRCUS-CD dataset in detail. Section III presents the proposed network AERNet. Section IV presents all the experimental results and a detailed discussion. In Section V, we summarize the work in this article and discuss future research directions.

II. HRCUS-CD Dataset

With the advancement of Earth observation technology over the past few decades, scholars have released open-sourced datasets for binary CD of RSIs, with gradually improved resolution. The release of these datasets has largely benefited the development of the CD domain [33]. We describe some datasets in Table I.

The SZTAKI Air Change Benchmark Set (SZTAKI) [15] has 13 pairs of 1.5-m spatial resolution optical aerial images of 952 × 640 pixels, which, in early research, was the earliest and most often used CD dataset. The Aerial Image Change Detection (AICD) [34] dataset is a simulated scene with 100 synthetic change aerial images at 0.5-m resolution, which has 1000 pairs of images (800 × 600 pixels) with major changes containing objects such as trees and buildings. The Onera Satellite Change Detection (OSCD) [35] dataset

TABLE I
PUBLIC DATASETS FOR BINARY CD IN RSIS

Datasets	Number of image pairs	Image size (pixels)	Resolution (m)
SZTAKI	13	952×640	1.5
AICD	1000	800×600	0.5
OSCD	24	600×600	10
SVCD	16000	256×256	0.03-1
WHU-building	1	32207×15354	0.2
LEVIR-CD	637	1024×1024	0.5
DSIFN	442	512×512	2
GDSCD	19	1006×1168-4936×5224	0.55
SYSU-CD	800	1024×1024	0.5
HRCUS-CD	11388	256×256	0.5

collects 24 pairs of Sentinel-2 multispectral satellite images taken between 2015 and 2018, each with a size of about 600 × 600 pixels and a resolution of 10 m. The Season-varying Change Detection (SVCD) [36] dataset has mainly two types of variations: synthetic images and real RSIs, and the commonly used real RSIs with seasonal variation contain 16 000 pairs of images of size 256 × 256 pixels at 0.03–1-m resolution. The WHU-building [37] CD dataset consists of a pair of aerial images gathered in 2012 (with 12 796 buildings) and 2016 (with 16 077 buildings), with a size of 32 207 × 15 354 pixels at 0.2-m resolution; the LEVIR-CD [21] dataset includes 637 pairs of RSIs of size 1024 × 1024 pixels with a resolution of 0.5 m obtained from Google Earth, and these two datasets are used very frequently in the BCD tasks. Images from the DSIFN [38] dataset are manually collected from Google Earth, and it contains 442 pairs of images (512 × 512 pixels) from six Chinese cities with a resolution of 2 m. Google Data Set for CD (GDSCD) [39] contains 19 pairs of satellite images with a resolution of 0.55 m, and the image pairs range in size from 1006 × 1168 pixels to 4936 × 5224 pixels and span the years 2006–2019. The Sun Yat-Sen University (SYSU-CD) [23] dataset consists of 800 pairs of images (1024 × 1024 pixels) at 0.5-m resolution, which were captured in Hong Kong between 2007 and 2014.

But some improvements can still be made. Some datasets prioritize suburban buildings and low-rise structures and may not adequately capture the changes in urban areas with complex environments, which pose greater challenges for detection. Furthermore, certain datasets have insufficient sample sizes for performing the BCD tasks. The proposed high-resolution complex urban scene BCD (HRCUS-CD) dataset largely complements the existing CD dataset in the above aspects. The dataset contains cropped 11 388 pairs of high-resolution RSIs with 256 × 256 pixels and at 0.5 m resolution, as well as more than 12 000 labeled change instances. All the labels are manually annotated by annotators with rich experience in RSIs' interpretation. This dataset was collected in Zhuhai, China, which has an area of 1736.45 km^2 and a resident population of approximately 2.44 million (as of November 2020). In recent years, Zhuhai has experienced rapid urbanization and industrialization.

The proposed HRCUS-CD dataset contains two main acquisition areas from two image sources: the first is mainly the urban built-up area, with a time span from 2019 to 2022. Considering the short time interval and the fact that this area is mostly built-up, the building changes' areas are small. The second area spans from 2010 to 2018, contains farmland and mountains, with a small number of old civil houses and buildings in the early period, and the area of building change is large later. These two types of high-resolution RSIs focus on built-up areas and new urban areas. The combination of these two areas leads to the strong diversity of our HRCUS-CD dataset.

The HRCUS-CD dataset proposed uses two satellite image sources to generate samples with fine annotation. This dataset boasts a large sample size and includes a wide range of complex environmental scenarios, such as urban villages, vegetation disturbances, high-rise apartments, and large contiguous buildings, including industrial parks and cultural and tourism facilities, as demonstrated in Fig. 1. The HRCUS-CD dataset integrates multiple time spans and multiple building change types, which improves the diversity and representativeness of the samples. It presents significant challenges for use in scientific research, providing more options and better benchmarks for evaluating BCD algorithms, promoting further advancements in the field. In addition, the HRCUS-CD dataset has practical applications for service needs such as illegal building detection and land resource management. Our developed HRCUS-CD dataset is freely accessible to all the researchers.

III. METHODOLOGY

A. Basic Network Structure

The proposed AERNet is a weight-sharing, two-branch, end-to-end network, as depicted in Fig. 2, which uses a classification-based approach. To transform dual-temporal images into a consistent feature space while preserving the distinct features of each individual image, the encoder component uses a weight-sharing feature extraction network (WFEN). The backbone of the WFEN is the architecture of the pretrained ResNet34 [40] before the global average pooling layer.

The decoder component is featured by a change discrimination network (CDN). After progressive abstraction of the convolution and pooling layers, the deepest features of the same scale extracted by WFEN1-5 and WFEN2-5 are stitched in the channel dimension and aggregated into the CDN. The aggregated features first enter the GCFAM to generate an initial change map with compact global information. The GCFAM can fully explore the long-range spatial–temporal dependencies between pixels (described in Section III-B). In the encoder, the resolution of the feature map is reduced to a very small size by the downsampling operations, which is not conducive to accurate segmentation. Given the crucial role that spatial domain information plays in the segmentation tasks, we stitch the features generated by the earlier layers of WFEN while passing them layer by layer through skip

Fig. 1. Examples of the HRCUS-CD dataset. (I) First region. (II) Second region. (a) Disappearing buildings. (b) New buildings. (c) New and disappeared. (d) and (e) Urban village. (f)–(h) Urban complex scenes.

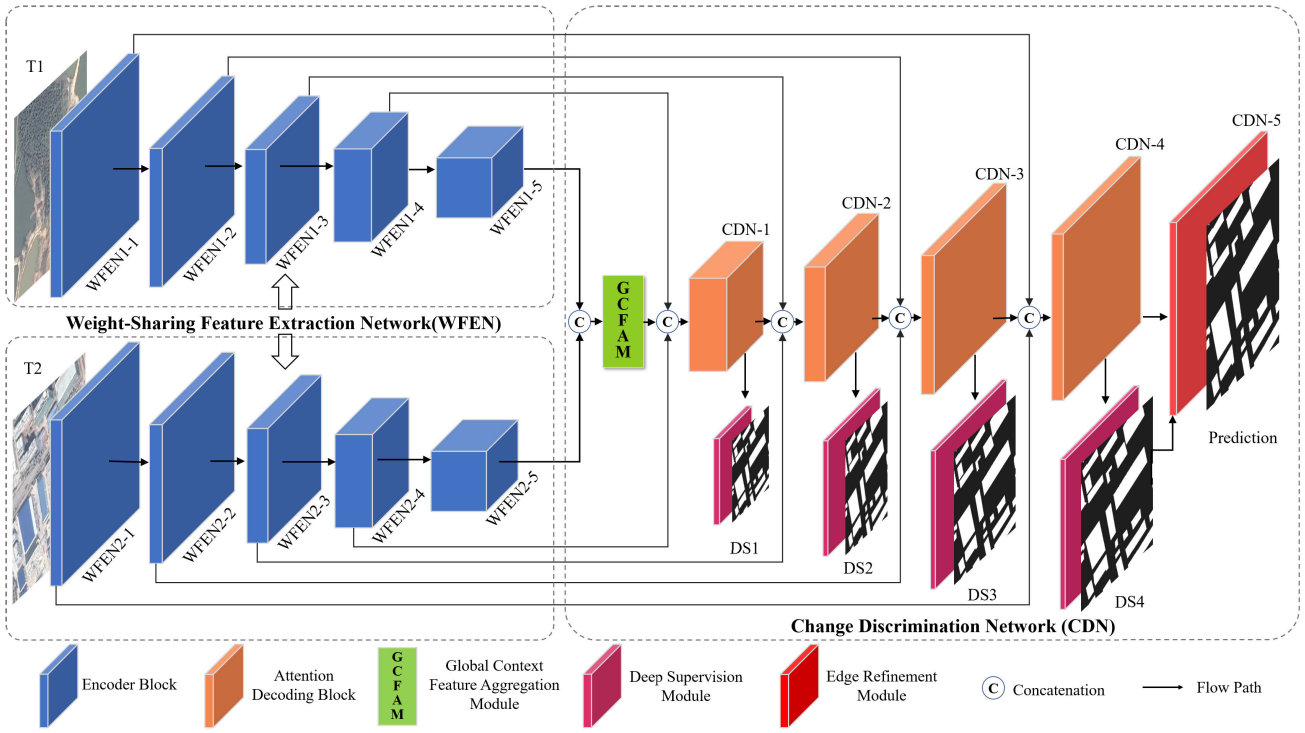


Fig. 2. Framework of the proposed AERNet.

connections, and then the features in series with GCFAM are fed into CDN-1, a lightweight attention decoding block (ADB) (described in Section III-C) that directs model's attention to changed buildings. Note that CDN-1–CDN-4 have the same structure. The features output from CDN-1, combined with the features transmitted by the corresponding skip connections, enter CDN-2–CDN-4 in turn, while the feature map's resolution continually recovers.

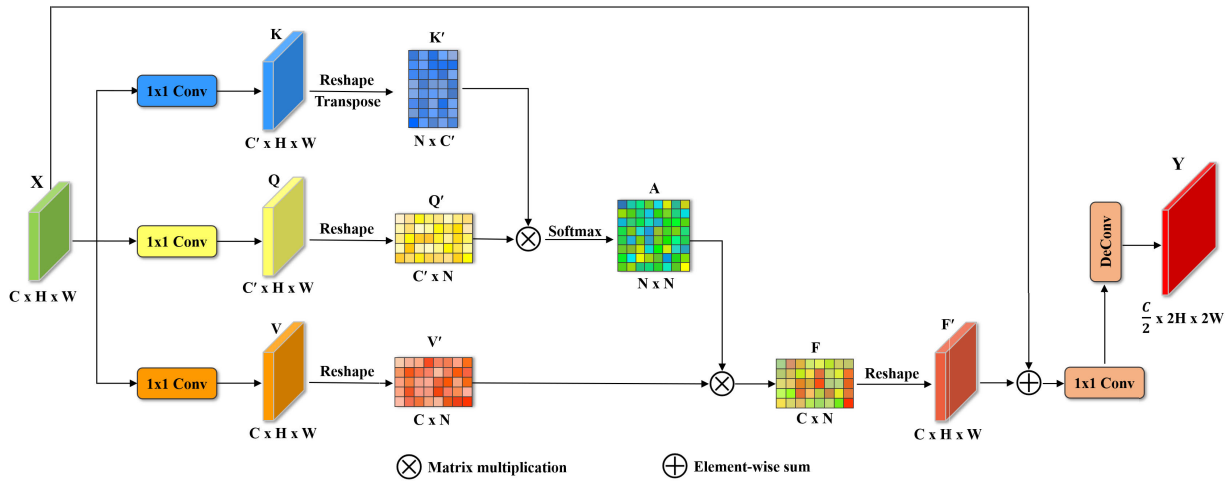


Fig. 3. Global context feature aggregation module.

In addition, the features in CDN-1–CDN-4 before upsampling are entered into DS1–DS4, respectively. DS facilitates model training and improves the network performance (described in Section III-E). The change maps generated by DS1–DS3 are used only to assist network training, while the change maps generated by DS4 are used to assist in training and the generation of final CDN-5. The features from CDN-4 match the original image's resolution, while change maps generated by DS4 need to be upsampled. Furthermore, the CDN-4 output features and the upsampled DS4 change maps are passed to CDN-5, an ERM where high-resolution representations with rich local details in different directions can be used to refine the edges (described in Section III-D).

B. Global Context Feature Aggregation Module

The global context feature aggregation module (GCFAM) used in this study bears a similar structure to the nonlocal neural network [41]. However, we made modifications to the GCFAM by adding a 1×1 convolution and transposed convolution for channel dimensionality reduction and upsampling at the end. It is an integral component of the AERNet and enables the modeling of the similarity between every pair of locations to extract the global context information dependency in features [42], as shown in Fig. 3.

First, the input features are defined as $X \in \mathbb{R}^{C\times H\times W}$, and then three 1×1 convolutions are used to convert X linearly to obtain three feature tensors: key ($K \in \mathbb{R}^{C'\times H\times W}$), query ($Q \in \mathbb{R}^{C'\times H\times W}$), and value ($V \in \mathbb{R}^{C\times H\times W}$), and C' represents the feature dimension of Q and K. Here, it is set to $C/8$. These three convolution operations bear the same structure, with the major difference lying in the output feature channels. We use different colors to represent the corresponding features. GCFAM reshapes and transposes K to obtain $K' \in \mathbb{R}^{N\times C'}$, reshapes Q to obtain $Q' \in \mathbb{R}^{C'\times N}$, and reshapes V to obtain $V' \in \mathbb{R}^{C\times N}$, where N is equal to the product of H and W. Then, GCFAM multiplies K' and Q' to obtain the initial attention map $A \in \mathbb{R}^{N\times N}$, and uses the Softmax function to map A to a range between 0 and 1 to model the long-distance dependence and similarity between pixel-level features

$$A = \text{Softmax}\left(K'Q'\right). \tag{1}$$

Then, multiply V' and A matrices to generate $F \in \mathbb{R}^{C\times N}$

$$F = V' \times A. \tag{2}$$

After reshaping F to obtain $F' \in \mathbb{R}^{C\times H\times W}$, the complete global context information is obtained.

Considering feature reuse and model convergence acceleration, F' and input X are summed pixel by pixelwise, and then the channels' number is reduced by 1×1 convolution so that the channels' number becomes half of the original. Finally, the final feature output $Y \in \mathbb{R}^{(C/2)\times 2W\times 2H}$ is obtained by deconvolution

$$Y = f^{\text{Deconv}}\left(f^{1\times 1}\left(X + F'\right)\right) \tag{3}$$

where f^{Deconv} denotes deconvolution for upsampling, and $f^{1\times 1}$ indicates 1×1 convolution to compress the channels. The derived feature Y captures the rich global context relationship between features.

C. Attention Decoding Block

Attention mechanisms, i.e., telling models "where" and "what" to pay attention to, have been frequently used to improve networks' performance [43], [44]. We use a more lightweight and effective coordinate attention (CA) [45], combined with depthwise separable convolution (DSConv) [46] to design the ADB (i.e., main structure from CDN-1 to CDN-4). CA consolidates features along two spatial directions, encoding global spatial information in a complementary fashion. This process captures remote dependencies between spatial locations that are essential for vision tasks, ultimately assisting the network in accurately locating objects of interest.

In ADB (Fig. 4), 1×1 convolution is first applied to integrate features across channels for the input features, i.e., dimensionality reduction. Then it goes to the core of ADB, which is our proposed enhanced CA (ECA) block. ECA consists of DSConv, CA, and residual connection. Initially, DSConv is used to extract features, followed by the utilization

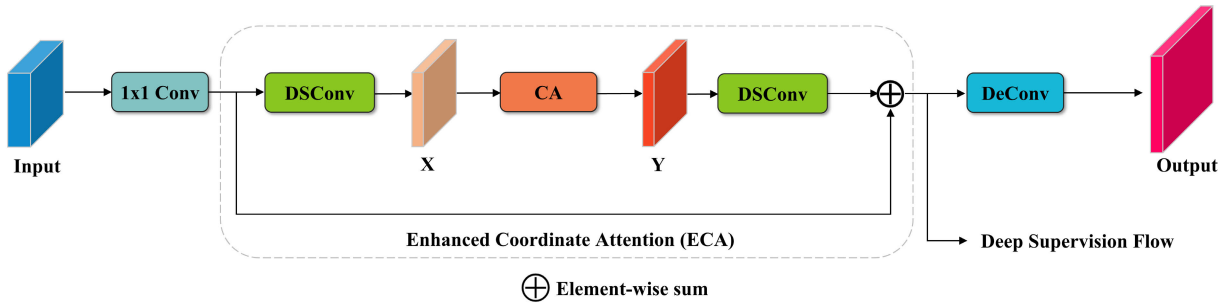


Fig. 4. Attention decoding block.

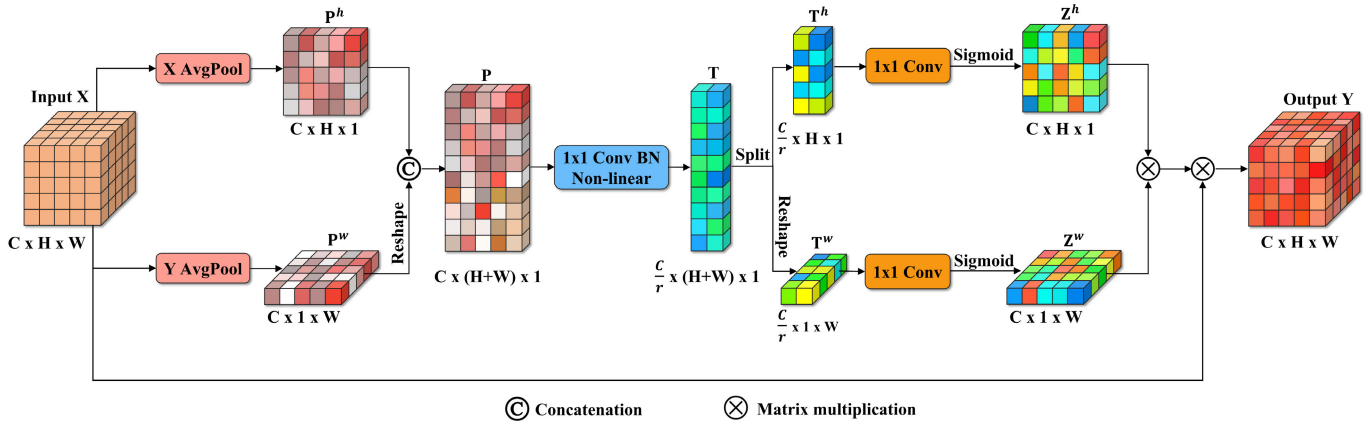


Fig. 5. Coordinate attention.

of CA to precisely locate changing buildings by leveraging its structural characteristics. CA emphasizes change information, thereby aiding network identification. Subsequently, DSConv is applied once more to optimize the features, which are then combined with residual connections to enhance the features further. ECA enables the AERNet to increase its network depth during the decoding stage while minimizing the number of parameters and computational overhead and simultaneously preventing gradient disappearance. By fully exploiting long-range dependencies and maximizing location information utilization, ECA facilitates more accurate identification of changing regions during the decoding stage. The features obtained via ECA are channeled through two separate pathways: one directed toward the DS module and the other toward the deconvolution (DeConv) block for upsampling. After four ADBs, the features' resolution is returned to its initial resolution.

Fig. 5 illustrates the specific framework of CA, with two primary steps. The first step is coordinate information generation: to enable the attention mechanism to record long-range spatial dependencies with precise location information, two 1-D pooling feature encoding operations are applied in CA to produce two feature maps with direction awareness. Specifically, the input features are $X = [x_1, x_2, \ldots, x_C] \in \mathbb{R}^{C\times H\times W}$, and to encode each channel of X along the horizontal and vertical directions, CA uses two spatial pooling kernels $(H, 1)$ and $(1, W)$. So, the cth channel's output at height h can be stated as

$$P_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h, i). \tag{4}$$

The cth channel's at width w can be stated as

$$P_c^w(w) = \frac{1}{H}\sum_{0\le j<H} x_c(j, w). \tag{5}$$

This pooling operation differs from the squeezing operation and allows the attention to retain exact location information in one spatial direction while capturing long-range interdependence in the other.

The second step is CA generation: CA applies a transformation to fully use the gathered location data so that the areas of interest can be effectively emphasized and the relationships between channels can be effectively captured. Specifically, given the features P^h and P^w generated by (4) and (5), CA first connects them in spatial dimensions to form $P \in \mathbb{R}^{C\times(H+W)\times 1}$, and then applies a 1×1 convolutional transformation $F_{1\times 1}$

$$T = \delta\left(F_{1\times 1}\left(\left[P^h, P^w\right]\right)\right) \tag{6}$$

where δ represents the nonlinear activation function, $[\cdot, \cdot]$ represents the join of spatial dimension, and $T \in \mathbb{R}^{(C/r)\times(H+W)\times 1}$ is the transformed spatial information feature map encoded in the horizontal and vertical directions, where r (set to 32 in

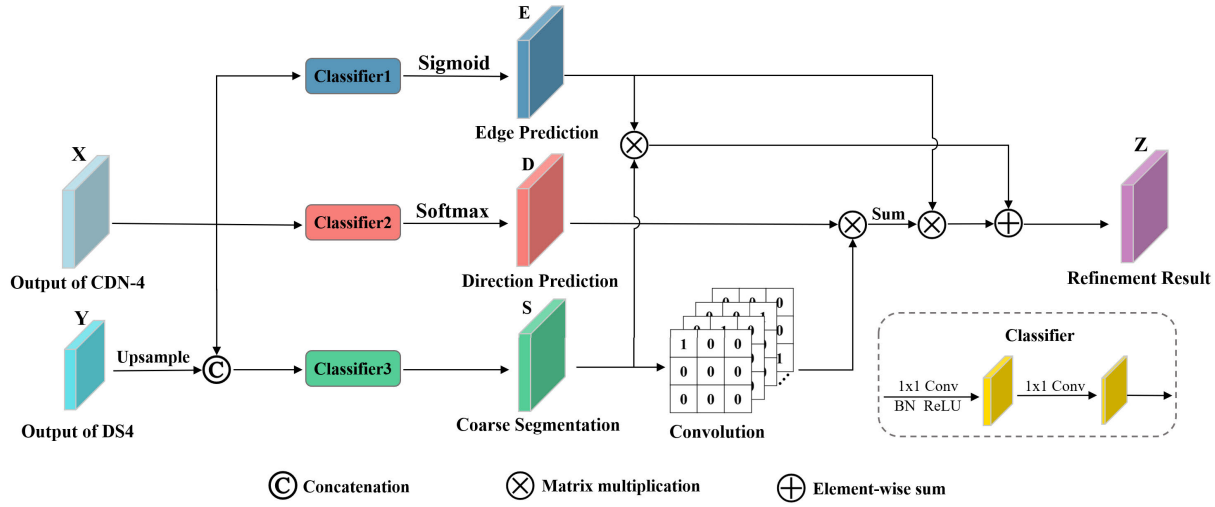


Fig. 6. Edge refinement module.

this study) is used to control the reduction rate of the feature block size.

Then CA splits T along the spatial dimension into two independent feature tensors $T^h \in \mathbb{R}^{(C/r)\times H\times 1}$ and $T^w \in \mathbb{R}^{(C/r)\times 1\times W}$, and two additional 1×1 convolution transforms $F^h_{1\times 1}$ and $F^w_{1\times 1}$ are used to transform T^h and T^w into a feature tensor equivalent to X in terms of channel counts. CA obtains

$$Z^h = \sigma\left(F^h_{1\times 1}\left(T^h\right)\right) \tag{7}$$

$$Z^w = \sigma\left(F^w_{1\times 1}\left(T^w\right)\right) \tag{8}$$

where σ represents the Sigmoid function, and the output feature of the final CA mechanism is $Y = \left[y_1, y_2, \ldots, y_C\right] \in \mathbb{R}^{C\times H\times W}$, denoted as

$$Y = Z^h \times Z^w \times X. \tag{9}$$

CA encodes spatial information, reassigning weights to different channels, and subsequently applies the final horizontal and vertical attention maps to input features through multiplication. This process allows CA to fully capitalize on the long-range dependencies between features and precise location information. The unique attention encoding approach empowers CA to accurately pinpoint the exact location of objects of interest, consequently generating more discernible features and enhancing the network's change recognition capability.

D. Edge Refinement Module

Since most buildings in RSIs often have regular boundaries, we introduce a plug-and-play, lightweight, and effective ERM [47], which updates edge pixels with high-resolution features and detailed local information.

As shown in Fig. 6, ERM takes the output X of CDN-4 and the prediction result Y of DS4 as inputs. First, to produce the edge prediction map E with a 0.5 threshold, X is fed into classifier1 and the Sigmoid function; meanwhile, X is fed into classifier2 and the Softmax function to generate the direction prediction map D; then Y is upsampled and connected with X in the channel dimension and fed into classifier3 to obtain the coarse segmentation prediction S

$$E = \sigma(F_{\text{classifier1}}(X)) > 0.5 \tag{10}$$

$$D = \rho(F_{\text{classifier2}}(X)) \tag{11}$$

$$S = F_{\text{classifier3}}([X; \text{up}(Y)]) \tag{12}$$

where σ represents the Sigmoid function, ρ represents the Softmax function, $up(\cdot)$ denotes the upsampling, $[\cdot;\cdot]$ denotes the feature splicing in the channel dimension, and $F_{\text{classifier}}$ denotes the classifier operation. As seen in the lower right portion of Fig. 6, these three classifiers have the same structure, with variations in their input and output features' channel counts. Each pixel's orientation to the nearest object's center is indicated by D. Eight directions are created by ERM by dividing 360° into 45° intervals, so classifier2 has eight channels of output features. D is represented as follows:

$$D^i = \frac{e^i}{\sum_j e^j} \tag{13}$$

where D^i represents the orientation prediction of pixel i; j represents eight orientations.

ERM replaces the pixel's predicted changed building confidence values with those of its direction prediction by combining D and S to obtain refined edges. Specifically, ERM designs a convolutional layer (Conv) to automatically refine the edges of changed buildings, which contains eight fixed unbiased convolutional kernels, the weights of which are frozen in training. Furthermore, the output features are aggregated along the channel with a summation operation to obtain an updated refinement region R

$$R = \text{sum}(\text{Conv}(S) \times D)_{\text{channel}}. \tag{14}$$

The goal of ERM is to refine only the edge regions. Thus, R is multiplied with E in the spatial dimension to enhance the edge regions, while the nonedge regions will be suppressed to keep the initial predicted results. Edge refinement results Z can be produced

$$Z = R \times E + S \times (1 - E). \tag{15}$$

E. DS and Self-Adaptive Weighted Loss Function

The weight parameters for CNN are optimized using the backpropagation algorithm. However, for networks with deeper depth and higher complexity, the supervision using just a loss function in the final output layer can easily lead to gradient disappearance during backpropagation, resulting in unstable parameter optimization in the network's middle layer, thus affecting the network performance [48].

To improve the aforementioned issues and boost the network's capacity for detection, in AERNet, we introduce the DS [49] strategy to ensure that the network's middle layer is properly supervised and discriminative for change region features. In total, there are four DS modules, of which the first three are only used to assist with training, and the last one passes the output prediction result to ERM while assisting with training. Specifically, for the change prediction map produced by each supervised branch, we calculate the loss by associating it with the label of the same spatial resolution obtained by downsampling the corresponding truth label

$$DS_i = \sigma\left(f^{1\times1}(ADB_i)\right) \tag{16}$$

where ADB_i denotes the feature branch of the ith CDN that flows into the DS module; $f^{1\times1}$ denotes the 1×1 convolution; and σ indicates the Sigmoid function.

We use the binary cross-entropy (BCE) loss function, which is often used in binary classification issues. However, in RSIs' CD tasks, the unchanged regions are usually much larger than the changed regions. Such an imbalance issue causes troubles for the model's attention [50]. Therefore, it is necessary to alleviate the effect of this imbalance, that is, to set weight coefficients on the loss function to constrain the network so that its training focuses more on the region of change. Based on BCE, we design a self-adaptive weighted BCE (SWBCE) loss function

$$L_{\text{SWBCE}} = -\frac{1}{m^*n}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[w_1 y_{(i,j)}\log\left(p_{(i,j)}\right) + w_2\left(1-y_{(i,j)}\right)\log\left(1-p_{(i,j)}\right)\right] \tag{17}$$

$$w_1 = \frac{\text{TN}}{\text{TN}+\text{FN}+\text{FP}} \tag{18}$$

$$w_2 = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{19}$$

where (i, j) denotes the pixel's spatial location index, $y_{(i,j)}$ denotes the label's value at point (i, j), generally 0 or 1, with 0 denoting no change, and 1 denoting change, $p_{(i,j)}$ represents the change map's value at point (i, j), m and n denote the map's dimensions, and w_1 and w_2 are the weights assigned. w_1 and w_2 correspond to the intersection over union (IoU) values of negative samples and positive samples, respectively. TP is the quantity of accurately identified changed pixels, and FP is the quantity of unchanged pixels mistakenly detected as changed pixels. TN indicates the amount of unchanged pixels detected correctly, and FN represents the amount of changed pixels mistakenly identified as unchanged pixels.

Note that both w_1 and w_2 are less than 1, and w_1's value is usually greater than w_2. To balance the network's attention, in the loss function, we assign w_1 and w_2 as the weights in the method of computation for changed and unchanged samples, respectively. In addition, w_1 and w_2 are updated in each epoch, corresponding to the loss calculation for each batch size, aiming to ensure that the loss function weights of each batch size correspond to the characteristics of the current samples. The integrated loss function L_{total} can be expressed as

$$L_{\text{total}} = \sum_{i=1}^{4}\lambda_i L^i_{\text{SWBCE}} + \lambda_{\text{out}} L^{\text{out}}_{\text{SWBCE}} \tag{20}$$

where λ_i and L^i_{SWBCE} denote the loss weights and loss values of the four supervision branches, respectively. λ_{out} and $L^{\text{out}}_{\text{SWBCE}}$ denote the weight of the ERM's final result of the network output in terms of the total loss and the loss value, respectively.

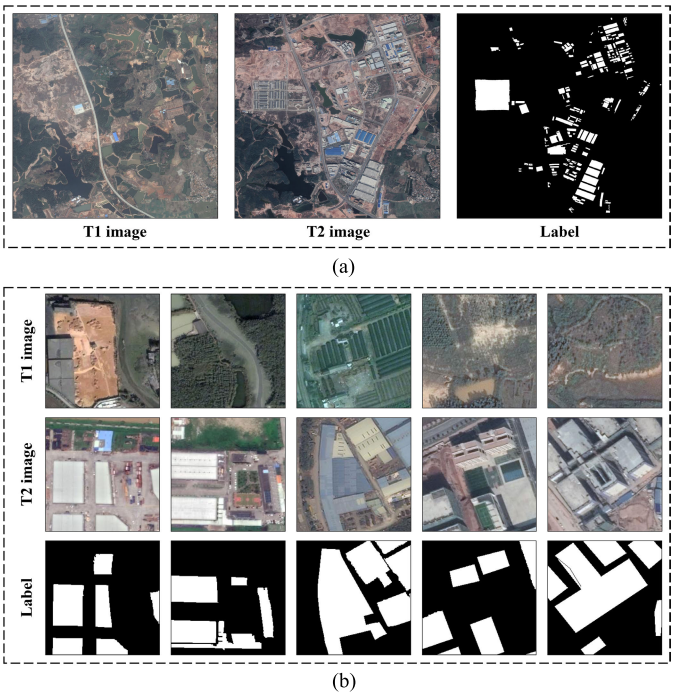


Fig. 7. Some selected examples of the GDSCD dataset. (a) Original images. (b) Cropped images.

IV. Experiment and Discussion

A. Datasets

Two publicly available datasets (i.e., GDSCD [39] and LEVIR-CD [21]) and the newly proposed HRCUS-CD dataset are used, which are described in Section II. Images in these datasets are cropped to 256×256 pixels in size, and the training set, validation set, and test set are allocated in different ratios: the GDSCD dataset contains 2883, 360, and 360 pairs of images, divided in an 8:1:1 ratio; the LEVIR-CD dataset includes 7120, 1024, and 2048 pairs of images, divided in a 7:1:2 ratio; and our HRCUS-CD dataset is divided in a ratio of 7:2:1 and contains 7974, 2276, and 1138 pairs of images. We use data enhancement methods that include rotation, transposition, flipping, and affine transformation. Selected examples of images of the GDSCD and LEVIR-CD datasets are shown in Figs. 7 and 8, respectively.

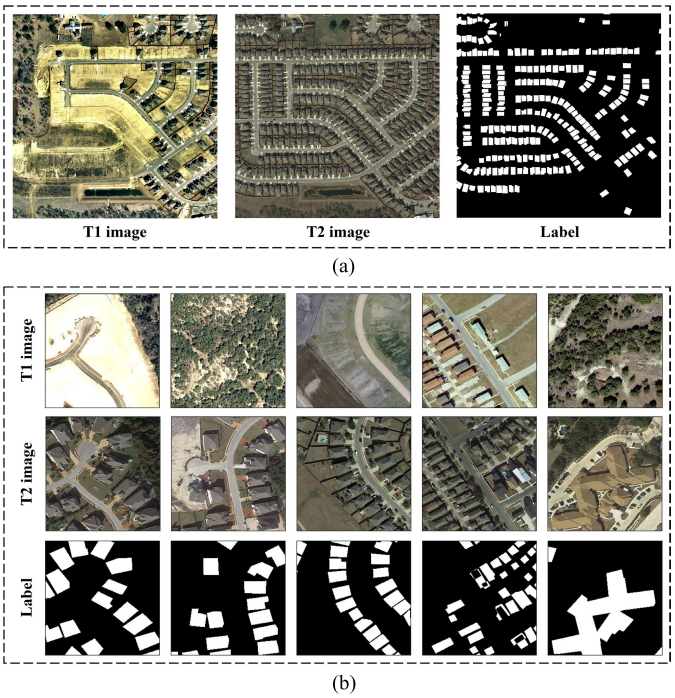


Fig. 8. Some selected examples of the LEVIR-CD dataset. (a) Original images. (b) Cropped images.

B. Competing Methods

To assess AERNet's performance, we compare it against the following 12 state-of-the-art (SOTA) CD methods, including FC-Siam-conc [51], FC-Siam-diff [51], UNet [52], SCDNet [53], ChangeNet [54], deeply supervised image fusion network (DSIFN) [38], cross-layer convolutional neural network (CLNet) [55], difference enhancement and spatial–spectral nonlocal network (DESSN) [56], bitemporal image transformer network (BIT) [57], ChangeFormer [30], intrascale cross-interaction and interscale feature fusion network (ICIFNet) [58], and dual-branch multilevel intertemporal network (DMINet) [59].

All the competing methods are trained from scratch. Considering that hyperparameters are quite important to model performance, to ensure fair comparison experiments, we follow the settings in their original literature.

C. Implementation Details

AERNet is implemented on the PyTorch deep learning framework and trained on NVIDIA TITAN XP with 12-GB memory. The optimizer used in the experiments is AdamW, the initial learning rate is set to 1e-4, the weight decay is 5e-4, CosineAnnealingWarmRestarts is used as the learning rate adjustment strategy, where the parameters T_0 is set to 4, T_{mult} is set to 2, the minimum learning rate eta_{min} is set to 1e-6, training is stopped after 125 epochs, and the batchsize is set to 24; for the proposed loss function [see (20)], λ_i and λ_{out} are set to 1.

In our experiments, precision (P), recall (R), $F1$-score ($F1$), and overall accuracy (OA) are used as evaluation criteria. Higher P means more accurate change pixels are detected, and higher R means the network's capacity to detect more change.

TABLE II
QUANTITATIVE RESULTS ON THE GDSCD DATASET. (THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD, THE SECOND-BEST PERFORMANCE IS EMPHASIZED BY UNDERLINE)

Methods	$P(\%)$	$R(\%)$	$F1(\%)$	$OA(\%)$	$1\_IoU(\%)$	$mIoU(\%)$
FC-Siam-conc	62.17	82.50	70.90	96.54	54.92	75.66
FC-Siam-diff	77.16	80.15	78.63	97.16	64.78	80.89
UNet	75.57	81.60	78.47	97.19	64.57	80.81
SCDNet	74.07	81.59	77.65	97.11	63.46	80.21
ChangeNet	79.67	<u>84.71</u>	82.11	97.65	69.65	83.58
DSIFN	78.47	83.88	81.09	97.52	68.19	82.78
CLNet	71.63	79.41	75.32	96.82	60.41	78.54
DESSN	76.36	81.88	79.02	97.25	65.32	81.21
BIT	<u>85.66</u>	74.66	79.78	97.44	66.37	81.83
ChangeFormer	81.68	71.41	76.20	96.98	61.55	79.19
ICIFNet	85.00	79.21	82.00	97.65	69.50	83.50
DMINet	85.37	79.88	<u>82.53</u>	<u>97.71</u>	<u>70.26</u>	<u>83.92</u>
AERNet	**88.74**	**89.18**	**88.96**	**98.51**	**80.12**	**89.27**

$F1$ represents a summed average of P and R, which is a metric for binary classification accuracy. OA is the ratio of pixels that were accurately categorized across all the categories

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{21}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{22}$$

$$F1 = \frac{2\text{PR}}{P + R} \tag{23}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{24}$$

Besides, unchanged region IoU (0_IoU), changed region IoU (1_IoU), and average IoU (mIoU) are also used as evaluation criteria. 0_IoU and 1_IoU mean the same as (18) and (19), and mIoU is the average of both. Larger values of these metrics indicate better performance of a model. The calculation of these indicators is as follows:

$$0\_\text{IoU} = \frac{\text{TN}}{\text{TN} + \text{FN} + \text{FP}} \tag{25}$$

$$1\_\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{26}$$

$$\text{mIoU} = \frac{0\_\text{IoU} + 1\_\text{IoU}}{2}. \tag{27}$$

TP, FP, TN, and FN have been described in Section III-E.

D. Experiments on the GDSCD Dataset

1) Quantitative Evaluation: As displayed in Table II, AERNet performs optimally: P (88.74%), R (89.18%), $F1$ (88.96%), OA (98.51%), 1_IoU (80.12%), and mIoU (89.27%). AERNet outperforms the optimal BIT model with an improvement of 3.08% in P, it outperforms the optimal ChangeNet model with an improvement of 4.47% in R, and it outperforms the best-performing competing method, DMINet, with improvements of 6.43%, 0.8%, 9.86%, and 5.35% for $F1$, OA, 1_IoU, and mIoU, respectively, on the GDSCD dataset.

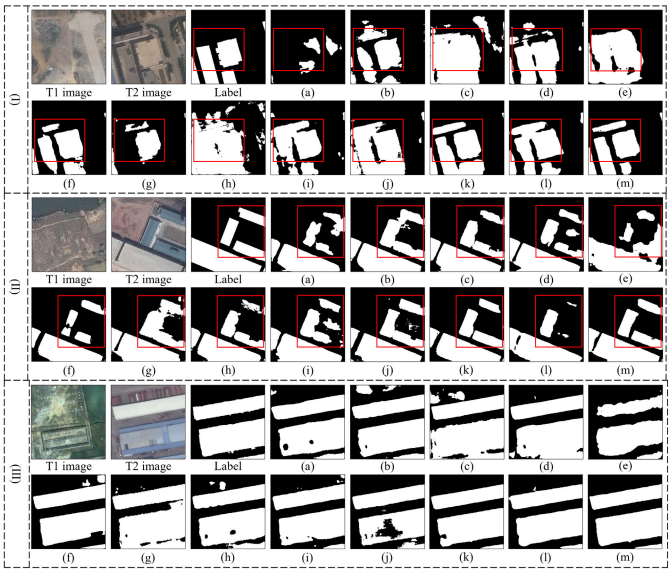


Fig. 9. Visual comparison of results on the GDSCD dataset. (a) FC-Siam–conc. (b) FC-Siam-diff. (c) UNet. (d) SCDNet. (e) ChangeNet. (f) DSIFN. (g) CLNet. (h) DESSN. (i) BIT. (j) ChangeFormer. (k) ICIFNet. (l) DMINet. (m) Our AERNet.

2) Qualitative Evaluation: Compared with other methods, AERNet's detection results (shown in Fig. 9) are the most visually similar to the ground-truth labels. AERNet produces complete and smoother building edges, resulting in fewer errors and omissions, and less adhesion between buildings. In addition, AERNet is capable of identifying changed buildings that were missed in the ground-truth labels, as highlighted by the red box in (I), and it shows great detection performance for small buildings [red box in (II)]. In addition, AERNet detects large buildings (III) with regular edges and no hollow leakage. DMINet has excellent quantitative and qualitative performance, and its detection results are the closest to AERNet among the compared methods. The high recall and average precision of ChangeNet are due to the fact that its qualitative results are mostly large adhesions, which often contain the region to be detected, resulting in a high error detection rate and a low miss detection rate.

E. Experiments on the LEVIR-CD Dataset

1) Quantitative Evaluation: As shown in Table III, AERNet achieves the highest $F1$, 1_IoU, and mIoU among all the evaluation metrics, with improvements of 0.79%, 1.31%, and 0.69%, respectively, compared with the best-performing competing method. It is worth noting that the LEVIR-CD dataset has higher quality samples, which results in considerably better performance for all the models. Despite this, AERNet still outperforms the SOTA models: BIT, ChangeFormer, and DMINet.

2) Qualitative Evaluation: Fig. 10 visualizes the three pairs of qualitative results from various methods. Visually, AERNet produces clear segmentation of the change region with few boundary adhesions, suppressing pseudochange and effectively avoiding the false identification of grasses [red box in (I)]. In addition, AERNet presents a strong performance

TABLE III
QUANTITATIVE RESULTS ON THE LEVIR-CD DATASET

Methods	$P(\%)$	$R(\%)$	$F1(\%)$	$OA(\%)$	$1\_IoU(\%)$	$mIoU(\%)$
FC-Siam-conc	83.02	87.47	85.19	98.53	74.20	86.33
FC-Siam-diff	83.86	89.69	86.68	98.69	76.49	87.56
UNet	85.08	90.48	87.70	98.78	78.09	88.41
SCDNet	84.99	88.99	86.94	98.70	76.90	87.77
ChangeNet	82.56	84.40	83.47	98.33	76.30	84.95
DSIFN	87.56	90.63	89.07	98.91	80.29	89.57
CLNet	87.94	90.49	89.20	98.92	80.50	89.68
DESSN	87.93	91.04	89.46	98.94	80.93	89.91
BIT	**92.85**	86.58	89.60	98.98	81.16	90.05
ChangeFormer	91.54	88.49	89.99	98.99	81.80	90.38
ICIFNet	90.66	88.47	89.55	98.95	81.08	89.99
DMINet	91.21	88.37	89.77	98.97	81.44	90.18
AERNet	89.97	**91.59**	**90.78**	**99.07**	**83.11**	**91.07**

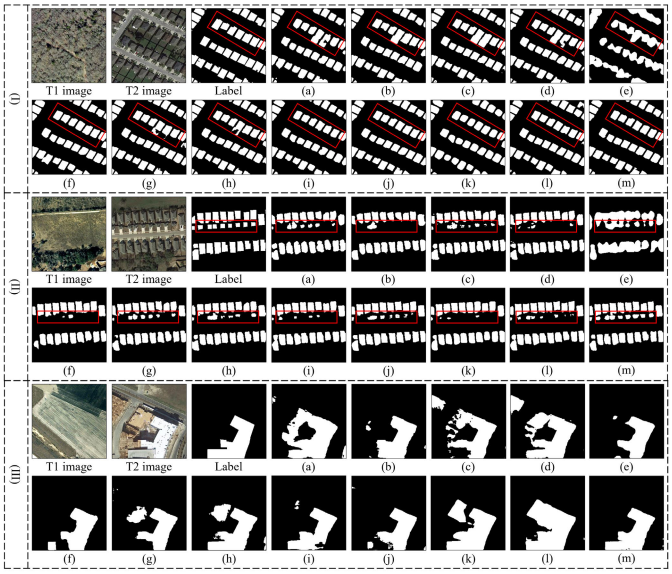


Fig. 10. Visual comparison of results on the LEVIR-CD dataset. (a) FC-Siam-conc. (b) FC-Siam-diff. (c) UNet. (d) SCDNet. (e) ChangeNet. (f) DSIFN. (g) CLNet. (h) DESSN. (i) BIT. (j) ChangeFormer. (k) ICIFNet. (l) DMINet. (m) Our AERNet.

in identifying very small change buildings [red box in (II)]. From the last pairs (III), mainly showing large irregular and complex buildings, AERNet can produce complete boundaries, which visually most closely matches the labels. While BIT, ChangeFormer, ICIFNet, and DMINet show promising results in detecting small buildings, their ability to detect large buildings falls behind that of AERNet, which explains why AERNet outperforms these methods in terms of evaluation metrics.

F. Experiments on HRCUS-CD Dataset

1) Quantitative Evaluation: As displayed in Table IV, AERNet achieves improvements of 7.0%, 8.78%, and 4.5% on $F1$, 1_IoU, and mIoU, respectively, compared with the metrics in the best-performing competing method. Note that the improvements from these three metrics are significant, with

TABLE IV
QUANTITATIVE RESULTS ON THE HRCUS-CD DATASET

Methods	$P(\%)$	$R(\%)$	$F1(\%)$	$OA(\%)$	$1\_IoU(\%)$	$mIoU(\%)$
FC-Siam-conc	53.95	66.95	59.75	98.64	42.61	70.61
FC-Siam-diff	64.29	67.76	65.98	98.76	49.23	73.99
UNet	60.87	69.75	65.01	98.77	48.16	73.46
SCDNet	55.74	67.95	61.24	98.68	44.14	71.40
ChangeNet	55.24	67.09	60.59	98.65	43.46	71.05
DSIFN	59.42	74.76	66.21	98.86	49.49	74.17
CLNet	67.67	61.07	64.20	98.59	47.28	72.92
DESSN	62.18	72.53	66.95	98.85	50.32	74.58
BIT	73.29	67.18	70.11	98.93	53.97	76.44
ChangeFormer	67.68	64.45	66.55	98.77	49.86	74.31
ICIFNet	70.00	60.42	64.86	98.77	47.80	73.38
DMINet	76.09	60.83	67.61	98.91	51.07	74.98
AERNet	**77.17**	**77.05**	**77.11**	**99.14**	**62.75**	**80.94**

TABLE V
QUANTITATIVE COMPARISON OF DIFFERENT METHODS WITHOUT/WITH SWBCE. [*REPRESENTS THE USE OF THE SWBCE LOSS FUNCTION, ALL THE SCORES ARE DESCRIBED IN PERCENTAGE (%)]

Methods	GDSCD			LEVIR-CD			HRCUS-CD		
	$F1$	1_IoU	$mIoU$	$F1$	1_IoU	$mIoU$	$F1$	1_IoU	$mIoU$
Baseline	85.36	74.46	86.17	89.37	80.78	89.84	72.95	57.42	78.23
Baseline*	**85.92**	**75.32**	**86.63**	**89.99**	**81.80**	**90.37**	**73.90**	**58.60**	**78.82**
AERNet	88.03	78.61	88.45	89.56	81.10	90.01	75.86	61.11	80.12
AERNet*	**88.96**	**80.12**	**89.27**	**90.78**	**83.11**	**91.07**	**77.11**	**62.75**	**80.94**
DSIFN	81.09	68.19	82.78	89.07	80.29	89.57	66.21	49.49	74.17
DSIFN*	**81.91**	**69.37**	**83.49**	**89.53**	**81.04**	**89.97**	**69.33**	**53.05**	**75.92**
CLNet	75.32	60.41	78.54	89.20	80.50	89.68	64.20	47.28	72.92
CLNet*	**76.58**	**62.04**	**79.40**	**89.47**	**80.95**	**89.92**	**65.32**	**48.50**	**73.54**
DESSN	79.02	65.32	81.21	89.46	80.93	89.91	66.95	50.32	74.58
DESSN*	**80.08**	**66.78**	**81.97**	**89.81**	**81.51**	**90.22**	**69.05**	**52.74**	**75.72**

two of them improving by more than 7%. In our developed HRCUS-CD dataset, AERNet demonstrates superior performance compared with other competing methods.

2) Qualitative Evaluation: The HRCUS-CD dataset is a challenging benchmark due to the complex environment surrounding the buildings. This is evident in Fig. 11, which presents three examples. In (I), AERNet achieves high accuracy in recognizing buildings in a complex surrounding scenario, with lower error and miss detection rates. In (II), urban villages with extremely complex environments are shown, and AERNet produces detection results that are visually similar to the labels. The last pair (III) contains uniformly colored factories with different sizes, as shown in the red box. Notably, competing methods fail to detect changes in small buildings or detect incomplete changes in large buildings. In comparison, the proposed AERNet mitigates these issues, as evidenced by the produced regular and complete boundaries.

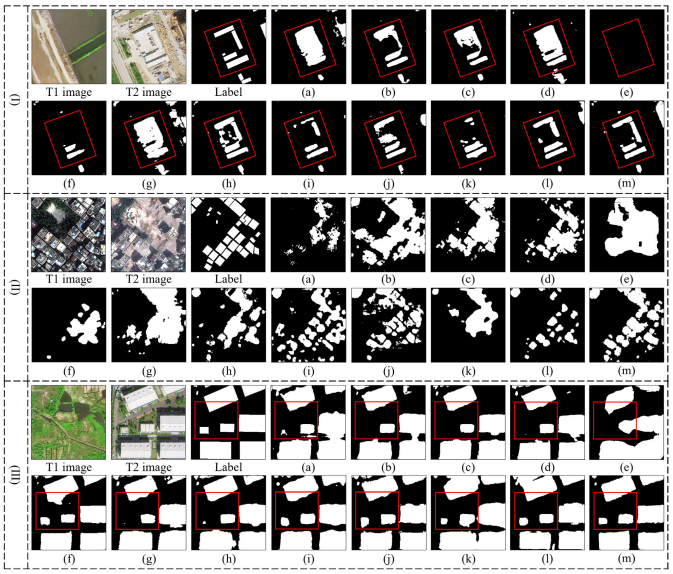


Fig. 11. Visual comparison of results on the HRCUS-CD dataset. (a) FC-Siam-conc. (b) FC-Siam-diff. (c) UNet. (d) SCDNet. (e) ChangeNet. (f) DSIFN. (g) CLNet. (h) DESSN. (i) BIT. (j) ChangeFormer. (k) ICIFNet. (l) DMINet. (m) Our AERNet.

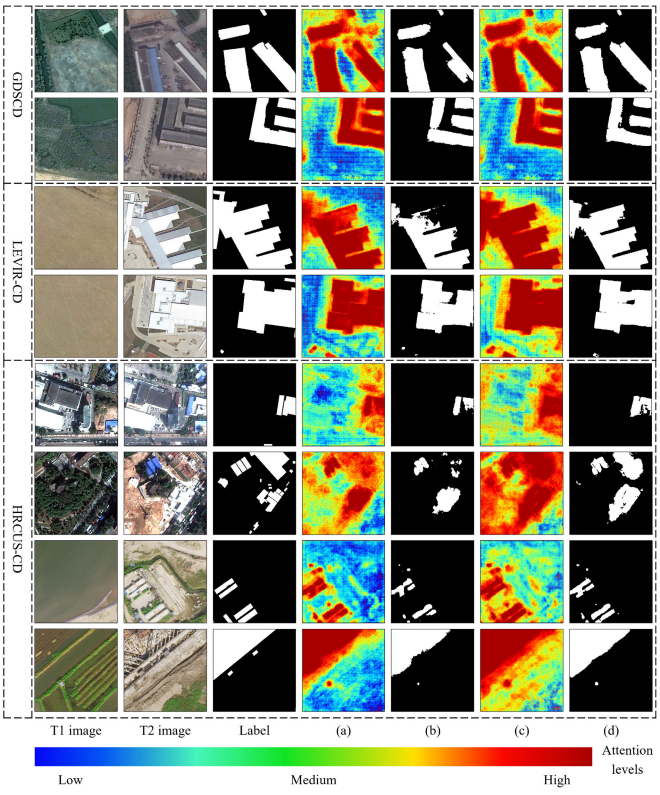


Fig. 12. Selected examples that demonstrate the impact of the SWBCE loss function on baseline. (a) Feature maps without SWBCE. (b) Detection results without SWBCE. (c) Feature maps with SWBCE. (d) Detection results with SWBCE.

G. Ablation Study of AERNet

1) Effect of the SWBCE Loss Function: To assess the effectiveness and generality of the SWBCE loss function, we conducted ablation study, including baseline (without any modules added), AERNet (with all the modules added), and three competing methods (DSIFN [38], CLNet [55], and DESSN [56]) using the BCE loss function. Table V shows the experimental results evaluated by $F1$, 1_IoU, and mIoU.

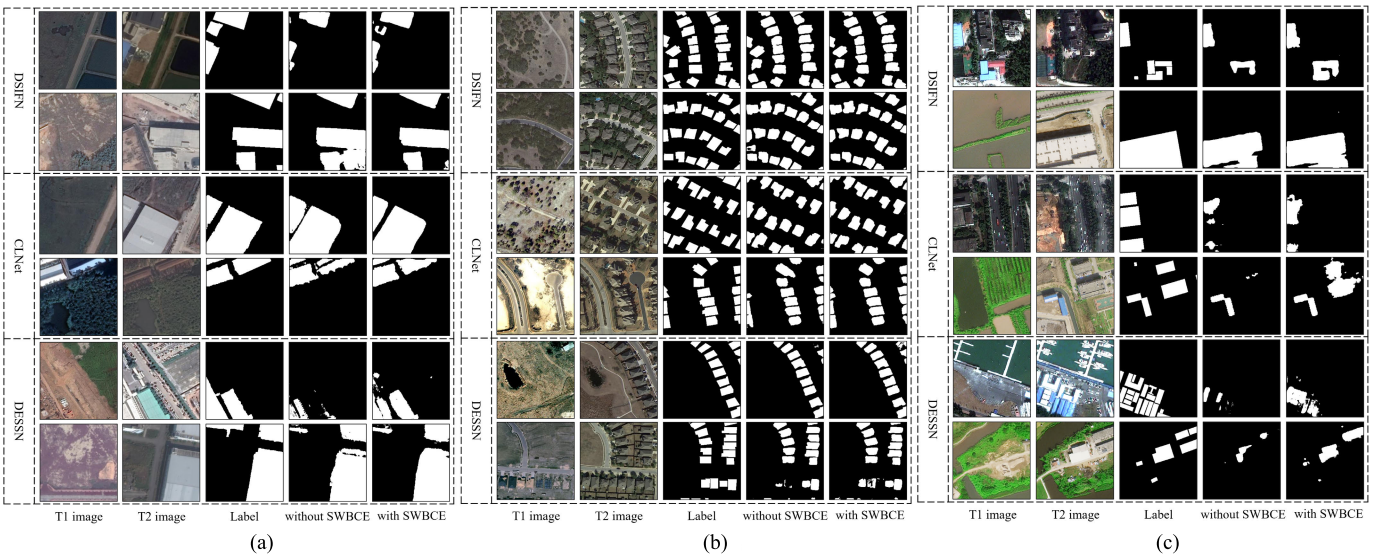


Fig. 13. Visual comparison of the impact of SWBCE loss function on the performance of three competing methods on three datasets. (a) GDSCD. (b) LEVIR-CD. (c) HRCUS-CD.

We observe that with the SWBCE loss function, the performance of both baseline and AERNet is improved, and this improvement is more pronounced on the fully structured AERNet. Moreover, DSIFN, CLNet, and DESSN showed improvements in their detection performance to varying degrees when using the SWBCE loss function.

Fig. 12 shows the visual feature maps and detection results without and with the SWBCE loss function, where the visual feature maps are from the output of the last layer of transposed convolution in CDN-4. It can be seen that the network pays significantly more attention to the changing building and its surrounding area and is more sensitive to changes. And the network positive detection rate is effectively increased, and the identified change subjects are more complete after using the SWBCE loss function on baseline, especially on the HRCUS-CD dataset. This demonstrates that the SWBCE loss function can improve the class imbalance and contribute to network performance optimization.

Fig. 13 visually demonstrates that three competing methods using the SWBCE loss function resulted in more complete detection of change regions, with fewer false detections and missed detections, leading to better overall detection results. Our experiments confirm the effectiveness and generosity of the SWBCE loss function, demonstrating that it can lead to improved performance in various methods. This further validates the importance of addressing the category imbalance problem and confirms that mitigating this problem can enhance the feature extraction capabilities of the network.

2) Ablation Study for Network Modules: To assess the effectiveness of GCFAM, ADB, ERM, and DS in AERNet, we conduct experiments where SWBCE is the loss function. From Table VI, we note that the main accuracy metrics show a generally increasing trend when modules are added. In addition, we note that the LEVIR-CD dataset's growing trend is unnoticeable. We believe that the LEVIR-CD dataset is easier to fit, and therefore the modules have limited effect on the overall improvement. For the GDSCD dataset with a small data volume and the HRCUS-CD dataset with complex imaging environments, our model components are clearly proven to benefit the improvement of BCD performance, the robustness of the model, and the generalization ability. The experiments on the three datasets prove that the modules we designed or used are effective.

TABLE VI

QUANTITATIVE PERFORMANCE COMPARISON OF ABLATION EXPERIMENTAL RESULTS ON DIFFERENT DATASETS

Dataset	Methods	GCFAM	ADB	ERM	DS	$F1(\%)$	$1\_IoU(\%)$	$mIoU(\%)$
GDSCD	Baseline+					85.92	75.32	86.63
	Baseline+	✓				86.74	76.58	87.33
	Baseline+	✓	✓			88.04	78.64	88.47
	Baseline+	✓	✓	✓		88.47	79.33	88.84
	Baseline+	✓	✓	✓	✓	**88.96**	**80.12**	**89.27**
LEVIR-CD	Baseline+					89.99	81.80	90.37
	Baseline+	✓				90.06	81.92	90.44
	Baseline+	✓	✓			90.32	82.35	90.67
	Baseline+	✓	✓	✓		90.42	82.52	90.75
	Baseline+	✓	✓	✓	✓	**90.78**	**83.11**	**91.07**
HRCUS-CD	Baseline+					73.90	58.60	78.82
	Baseline+	✓				74.81	59.76	79.40
	Baseline+	✓	✓			75.27	60.34	79.68
	Baseline+	✓	✓	✓		76.11	61.43	80.26
	Baseline+	✓	✓	✓	✓	**77.11**	**62.75**	**80.94**

H. Comparison of Efficiency

In the efficiency comparison, we use the number of parameters (Params) and floating-point operations (FLOPs) to assess the efficiency of various methods. The size of the images used for efficiency comparison is 256 × 256 pixels. The FLOPs for each method are represented as the mean value needed to process a pair of images, while Params are constant. Table VII reveals the first two networks, with smaller Params

TABLE VII
EFFICIENCY COMPARISON OF DIFFERENT METHODS

Methods	Params(M)	FLOPs(G)
FC-Siam-conc	1.54	5.29
FC-Siam-diff	1.35	4.74
UNet	31.04	54.85
SCDNet	2.03	40.45
ChangeNet	47.20	10.91
DSIFN	35.73	82.26
CLNet	8.10	8.67
DESSN	19.35	36.81
BIT	3.50	10.63
ChangeFormer	41.03	202.79
ICIFNet	23.84	24.51
DMINet	6.24	14.55
AERNet	25.36	12.82

TABLE VIII
QUANTITATIVE RESULTS ON THE WHU-BUILDING DATASET. (THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD, THE SECOND-BEST PERFORMANCE IS EMPHASIZED BY UNDERLINE)

Methods	$P(\%)$	$R(\%)$	$F1(\%)$	$OA(\%)$	$1\_IoU(\%)$	$mIoU(\%)$
FC-Siam-conc	86.51	75.15	80.43	98.48	67.27	82.85
FC-Siam-diff	82.09	86.03	84.01	98.87	72.43	85.63
UNet	85.87	88.57	87.20	99.09	77.30	88.18
SCDNet	82.93	86.08	84.48	98.90	73.12	85.99
ChangeNet	78.40	<u>89.80</u>	83.71	98.90	71.99	85.43
DSIFN	92.77	87.24	<u>89.92</u>	99.25	<u>81.69</u>	<u>90.45</u>
CLNet	88.59	89.68	89.13	99.22	80.40	89.79
DESSN	87.80	88.01	87.91	99.13	78.42	88.76
BIT	88.22	85.14	86.66	99.05	76.45	87.74
ChangeFormer	89.60	85.29	87.39	99.11	77.61	88.34
ICIFNet	<u>92.81</u>	86.53	89.56	<u>99.27</u>	81.10	90.17
DMINet	90.17	89.14	89.65	99.26	81.25	90.24
AERNet	**92.92**	**90.02**	**91.45**	**99.37**	**84.24**	**91.79**

and FLOPs, but they exhibit the general CD capability, mainly due to the reduced number of channels in their decoder components. ChangeNet has the largest Params (47.20 M), while DSIFN has 35.73 M Params and the second-largest FLOPs (82.26G) after ChangeFormer, which has 41.03 M Params and the largest FLOPs (202.79G). CLNet has lower Params and FLOPs, while DESSN has achieved a certain balance in Params and FLOPs. BIT shows good detection performance with lower Params and FLOPs. ICIFNet has moderate Params and FLOPs and good robustness; DMINet has lower Params and FLOPs and good detection performance. Although AERNet has slightly higher Params, its FLOPs are lower, only 12.82G, which is similar to BIT and much lower than ChangeFormer. Furthermore, AERNet has the best CD performance with good robustness and generalization compared with other methods, indicating that it achieves a great balance in terms of Params, FLOPs, and performance.

V. CONCLUSION

In this article, we propose an AERNet for BCD, which consists of WFEN and CDN. AERNet uses the GCFAM to aggregate the extracted multilayer contextual information and enhance the relationship between global features. We introduce a lightweight ADB guided by ECA; it enhances the focus on pixels of interest and the ability to discriminate between change features in both the channel and spatial dimensions. The ERM is used to improve the network's ability to sense and refine the changing edges of buildings. To better guide the learning process of the network, we combine the newly designed SWBCE loss function with the DS strategy to both alleviate sample category imbalance and optimize capacity of the network to fit features.

What is more, the experiments on GDSCD, LEVIR-CD, and our newly developed HRCUS-CD datasets show that the proposed AERNet outperforms other SOTA methods, demonstrating strong robustness and superior generalization performance. The results of ablation studies show that the model components we use or design are effective for AERNet, and the proposed SWBCE loss function has good generalizability. The efficiency comparison shows that AERNet achieves optimal performance while achieving a great balance between Params and FLOPs. Future research will focus on weakly supervised semantic CD methods.

APPENDIX

To further validate the detection performance of the AERNet, we conducted a comparison test on the WHU-building dataset [37], which has a resolution of 0.2 m. The images in the WHU-building dataset are cropped to a size of 256 × 256 pixels. The dataset consists of 5946 training pairs, 744 validation pairs, and 744 testing pairs, with a ratio of 8:1:1.

A. Quantitative Evaluation

According to Table VIII, AERNet outperforms all other methods in all the evaluation metrics, with $F1$, 1_IoU, and mIoU improving by 1.53%, 2.55%, and 1.34%, respectively, compared with the best-performing competing methods. BIT and ChangeFormer perform averagely, and the best-performing competitive method is DSIFN, followed by ICIFNet and DMINet.

B. Qualitative Evaluation

Fig. 14 illustrates the qualitative results of the investigated methods. Visually, AERNet achieves the highest completeness of detection results, the lowest void miss detection rate, and best matches the real change results, indicating that AERNet is able to effectively extract global information and maximize the identification of global change features. In particular, for a building with very similar ground color (I), AERNet is able to suppress pseudochange and identify the change area completely. For a dense small building (II), AERNet produces results with clear edges and complete main body, with an extremely low false detection rate. For the changing building (III) with a complex background (vehicles, roads,

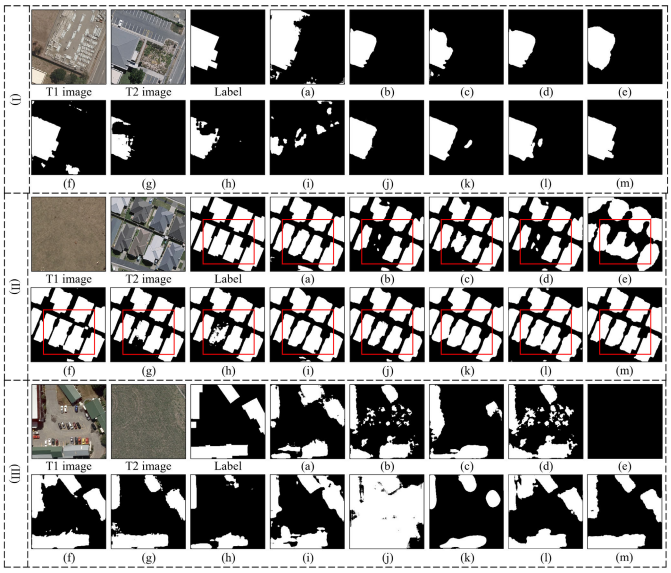


Fig. 14. Visual comparison of results on the WHU-building dataset. (a) FC-Siam-conc. (b) FC-Siam-diff. (c) UNet. (d) SCDNet. (e) ChangeNet. (f) DSIFN. (g) CLNet. (h) DESSN. (i) BIT. (j) ChangeFormer. (k) ICIFNet. (l) DMINet. (m) Our AERNet.

and shadows), AERNet achieves a complete segmentation result, effectively filtering out the interference of irrelevant factors. These results demonstrate that AERNet has superior generalization performance and robustness, even in the face of higher resolution images.

REFERENCES

[1] S. Xu and S. Fina, "National-scale imperviousness mapping and detection of urban land changes," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 369–384, Aug. 2023.

[2] C. Peng et al., "Global spatiotemporal trend of satellite-based soil moisture and its influencing factors in the early 21st century," *Remote Sens. Environ.*, vol. 291, Jun. 2023, Art. no. 113569.

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

[4] Y. Lu, N. C. Coops, and T. Hermosilla, "Estimating urban vegetation fraction across 25 cities in pan-Pacific using Landsat time series data," *ISPRS J. Photogramm. Remote Sens.*, vol. 126, pp. 11–23, Apr. 2017.

[5] H. Xu, S. Qi, X. Li, C. Gao, Y. Wei, and C. Liu, "Monitoring three-decade dynamics of citrus planting in Southeastern China using dense Landsat records," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102518.

[6] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2403–2420, May 2010.

[7] M. Awrangjeb, S. Gilani, and F. Siddiqui, "An effective data-driven method for 3-D building roof reconstruction and robust change detection," *Remote Sens.*, vol. 10, no. 10, p. 1512, Sep. 2018.

[8] M. Gong, P. Zhang, L. Su, and J. Liu, "Coupled dictionary learning for change detection from multisource data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7077–7091, Dec. 2016.

[9] T. Bai et al., "Deep learning for change detection in remote sensing: A review," *Geo-Spatial Inf. Sci.*, pp. 1–27, 2022, doi: 10.1080/10095020.2022.2085633.

[10] L. Ke, Y. Lin, Z. Zeng, L. Zhang, and L. Meng, "Adaptive change detection with significance test," *IEEE Access*, vol. 6, pp. 27442–27450, 2018.

[11] S. Ye, D. Chen, and J. Yu, "A targeted change-detection procedure by combining change vector analysis and post-classification approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 115–124, Apr. 2016.

[12] H. Zhuang, K. Deng, H. Fan, and M. Yu, "Strategies combining spectral angle mapper and change vector analysis to unsupervised change detection in multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 681–685, May 2016.

[13] 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.

[14] Y. Zhang, D. Peng, and X. Huang, "Object-based change detection for VHR images based on multiscale uncertainty analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 13–17, Jan. 2018.

[15] C. Benedek and T. Sziranyi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.

[16] J. Zheng and H. You, "A new model-independent method for change detection in multitemporal SAR images based on radon transform and Jeffrey divergence," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 91–95, Jan. 2013.

[17] L. Song, M. Xia, J. Jin, M. Qian, and Y. Zhang, "SUACDNet: Attentional change detection network based on Siamese U-shaped structure," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102597.

[18] Q. Zhu, X. Guo, Z. Li, and D. Li, "A review of multi-class change detection for satellite remote sensing imagery," *Geo-Spatial Inf. Sci.*, pp. 1–15, 2022, doi: 10.1080/10095020.2022.2128902.

[19] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.

[20] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodríguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.

[21] 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, p. 1662, May 2020.

[22] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[23] 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, 2022, Art. no. 5604816.

[24] M. Wang, K. Tan, X. Jia, X. Wang, and Y. Chen, "A deep Siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images," *Remote Sens.*, vol. 12, no. 2, p. 205, Jan. 2020.

[25] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

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

[27] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.

[28] Q. Ding, Z. Shao, X. Huang, and O. Altan, "DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102591.

[29] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," 2021, *arXiv:2108.07002*.

[30] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," 2022, *arXiv:2201.01293*.

[31] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[32] S. Gao, W. Li, K. Sun, J. Wei, Y. Chen, and X. Wang, "Built-up area change detection using multi-task network with object-level refinement," *Remote Sens.*, vol. 14, no. 4, p. 957, Feb. 2022.

[33] S. Tian, Y. Zhong, Z. Zheng, A. Ma, X. Tan, and L. Zhang, "Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application," *ISPRS J. Photogramm. Remote Sens.*, vol. 193, pp. 164–186, Nov. 2022.

[34] N. Bourdis, D. Marraud, and H. Sahbi, "Constrained optical flow for aerial image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 4176–4179.

[35] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," 2018, *arXiv:1810.08468*.

[36] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, May 2018.

[37] 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.

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

[39] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[42] D. Feng et al., "GCCINet: Global feature capture and cross-layer information interaction network for building extraction from remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 114, Nov. 2022, Art. no. 103046.

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[44] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10093–10102.

[45] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[46] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[47] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 240–252, Jan. 2022.

[48] J. Liu, M. Gong, A. K. Qin, and K. C. Tan, "Bipartite differential neural network for unsupervised image change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 876–890, Mar. 2020.

[49] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, vol. 38, G. Lebanon and S. V. N. Vishwanathan, Eds., San Diego, CA, USA, May 2015, pp. 562–570.

[50] T. Leichtle, C. Geiß, T. Lakes, and H. Taubenböck, "Class imbalance in unsupervised change detection—A diagnostic analysis from urban remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 60, pp. 83–98, Aug. 2017.

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

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.

[53] P. Fernández Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," in *Proc. Robot., Sci. Sys. (RSS)*, Jun. 2016, pp. 1–16.

[54] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCV)*, Sep. 2018, pp. 129–145.

[55] Z. Zheng, Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang, "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 247–267, May 2021.

[56] T. Lei et al., "Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4507013.

[57] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416.

[58] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3168331.

[59] 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, 2023, Art. no. 4401015.

Jindou Zhang received the B.S. degree in surveying and mapping engineering from Chongqing University, Chongqing, China, in 2021. He is currently pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

His research interests include deep learning, remote sensing change detection, and object detection.

Zhenfeng Shao received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2004.

Since 2009, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He has authored or coauthored more than 50 peer-reviewed articles in international journals. His research interests include high-resolution image processing, pattern recognition, and urban remote sensing applications.

Dr. Shao was a recipient of the Talbert Abrams Award for the Best Paper in Image Matching from the American Society for Photogrammetry and Remote Sensing in 2014 and the New Century Excellent Talents in University from the Ministry of Education of China in 2012. Since 2019, he has been serving as an Associate Editor for the *Photogrammetric Engineering & Remote Sensing* (PE & RS) specializing in smart cities, photogrammetry, and change detection.

Qing Ding received the bachelor's degree in surveying and mapping and the master's degree in cartography and geographic information engineering from Jilin University, Changchun, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

His research interests include urban remote sensing mapping and change detection.

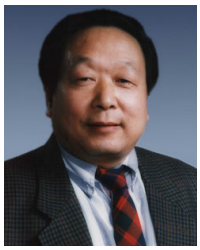
Xiao Huang is an Assistant Professor with the Department of Geosciences, University of Arkansas, Fayetteville, AR, USA. His research interests include GeoAI, remote sensing, spatial machine learning, social sensing, and disaster mitigation.

Yu Wang received the master's degree from the Wuhan Institute of Technology, Wuhan, China, in 2021. He is currently pursuing the Ph.D. degree with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, under the supervision of Prof. Zhenfeng Shao.

His research interests include image/video processing and computer vision.

Xuechao Zhou received the B.S. degree in geographic information science from Central South University, Changsha, China, in 2021. He is currently pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

His research interests include high spatial resolution remote sensing image classification, remote sensing applications, and image processing.



Deren Li received the Ph.D. degree in photogrammetry from the University of Stuttgart, Stuttgart, Germany, in 1986, and the Honorary Ph.D. degree from ETH Zürich, Zürich, Switzerland, in 2008.

He is a Scientist in surveying, mapping, and remote sensing with Wuhan University, Wuhan, China.

Dr. Li is a member of the Chinese Academy of Sciences and the Chinese Academy of Engineering. He is also a member of the International Eurasia Academy of Sciences and International Academy of Astronautics. He was a recipient of the Honorary Member and the Brock Gold Medal in recognition of outstanding contributions to photogrammetry from the International Society for Photogrammetry and Remote Sensing.