Remote Sensing Change Detection via Centralized Low-High Rank Representation

Yikui Zhai, *Senior Member, IEEE,* Zhanyu Shen, Junying Zeng, *Member, IEEE,* Hongsheng Zhang, *Senior Member, IEEE,* Hufei Zhu, *Member, IEEE,* Qingling Chang, *Member, IEEE,* Yan Cui, *Member, IEEE,* Pasquale Coscia, *Senior Member, IEEE,* Angelo Genovese, *Senior Member, IEEE*

***Abstract*—Recently, the advancement of remote sensing equipment and image processing technology has resulted in high-resolution remote sensing (HRRS) images containing more detailed geospatial characteristic. This progress has increased the demands on both the recognition capabilities and the detection speed required for remote sensing change detection (RSCD). Due of a wide variety of surface objects on the earth, RSCD tasks are characterized by "intra-class variance" and "inter-class similarity". Current mainstream lightweight deep learning models often use fewer network layers to maintain a low model parameters. Therefore, lightweight networks often struggle to capture the rich contextual information present in HRRS images, making comprehensive change detection challenging, specifically as shallow networks struggle to maintain contextual information of single-temporal images while fully detecting changes in bitemporal images. In the paper, we introduce a straightforward and lightweight network called Centralized High-Low Rank and Fusion Network (CHLF-Net), which primarily consists of two modules. First, we design a Centralized High-Low Rank Representation (CHLR) module to extract intra-temporal and inter-temporal information in two ways. One is High-Rank Representation Aggregation (HRRA), which uses an MLP structure to compute difference information between bitemporal images; the other is Low-Rank Representation Comparison (LRRC), which uses learnable low-rank matrices to capture contextual information of single-temporal images and compares the obtained bitemporal low-rank matrices to acquire highly discriminative change features. Second, we design a Lightweight Deep-Shallow Attention (LDSA) module, which lightweightly fuses deep and shallow features, enabling a better integration of rich contextual information and detailed multi-scale features. Employing a shallow backbone network without intricate structures, CHLF-Net, with just 1.27M model parameters, surpasses other advanced methods on three RSCD datasets, showcasing its superiority. Our code is available at https://github.com/yikuizhai/CHLF-Net.**

***Index Terms*—Change detection (CD), Centralized High-Low Rank Representation (CHLR), Lightweight Deep-Shallow Fusion Attention (LDSA).**

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011576; in part by Guangdong Higher Education Innovation and Strengthening School Project under Grant 2020ZDZX3031, Grant 2022ZDZX1032, Grant 2023ZDZX1029, and Grant 2024ZDZX1009; in part by the Wuyi University Hong Kong and Macao Joint Research and Development Fund under Grant 2022WGALH19; and in part by Guangdong Jiangmen Science and Technology Research Project under Grant 2220002000246 and Grant 2023760300070008390. (Yikui Zhai, Zhanyu Shen, Junying Zeng, Hufei Zhu, Qingling Chang and Yan Cui contributed equally to this work.) (Corresponding authors: Yikui Zhai; Hongsheng Zhang.)

Yikui Zhai, Zhanyu Shen, Junying Zeng, Hufei Zhu and Qingling Chang are with the School of Electronics and lnformation Engineering, Wuyi University, Jiangmen, 529020, China (e-mail: yikuizhai@163.com; zhanyushen.s@gmail.com; zengjunying@126.com; hufeizhu93@hotmail.com; qlchangcas@gmail.com).

Hongsheng Zhang is with the Department of Geography, The University of Hong Kong, Hong Kong, China (e-mail: zhanghs@hku.hk).

Yan Cui are with the School of Electronics and lnformation Engineering, Wuyi University, Jiangmen, 529020, China, and also with with Zhuhai 4Dage Network Technology, Zhuhai 519000, China (e-mail: cuiyan@wyu.edu.cn).

Pasquale Coscia, Angelo Genovese are with the Dipartimento di Computer Science and the Dipartimento di Informatica, Universita degli Studi di Milano, 20133 Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it).

Manuscript completed November 10, 2024.

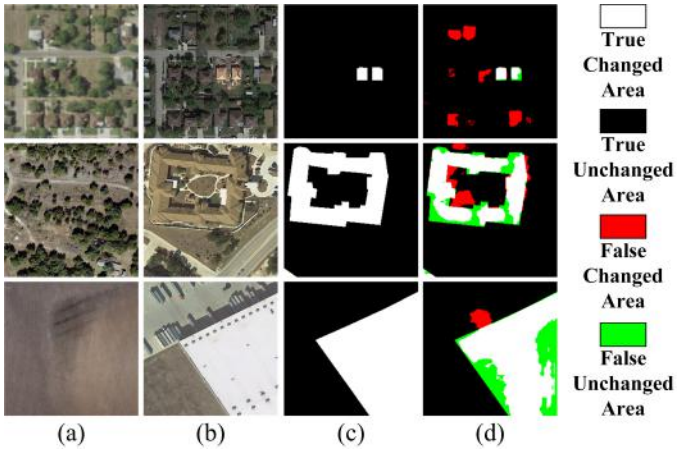


Fig. 1. Illustrate the importance of capturing high-level semantics of targets through three pairs of bitemporal HRRS images: drastic spectral changes cause false alarms (in the first row); the foreground and the background having similar spectral information leads to missed detections (in the second row); failed to capture the complete semantics of large-scale change areas (in the third row). The images are arranged as follows: (a) T1 image, (b) T2 image, (c) ground truth, and (d) potential prediction change map.

I. Introduction

REMOTE sensing change detection (RSCD) focuses on segmenting fine regions between change and no-change categories or distinguishing various semantic classes of change areas in bitemporal images, captured at different times but at the same location from an aerial view using satellites, spacecraft, or drones. In recent years, RSCD has become a significant focus of theoretical research in Earth remote sensing observation and has become essential in numerous practical applications, such as monitoring surface resources [1], disaster monitoring [2], [3] and urban inspection [4], [5].

Remote sensing sensors and Earth observation technology's progression has enriched high-resolution remote sensing (HRRS) images with more spatial information. This progress offers new opportunities for large-scale land cover and use

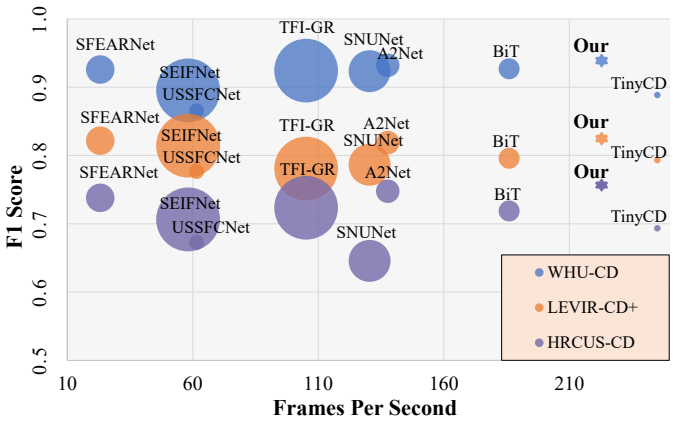


Fig. 2. Compare model performance and complexity across three datasets: HRCUS-CD, WHU-CD, and LEVIR-CD+. The comparison of Nine methods includes model parameters (Params), computational cost (FPS), and detection accuracy (F1 score). The rendered colors represent HRCUS-CD (perple), WHU-CD (blue), and LEVIR-CD+ (orange). For brevity, circle size is used to represent Params; the larger the circle, the larger the parameters, and vice versa. The proposed method, CHLF-Net, achieves higher accuracy with faster detection speed and smaller model parameters.

technology development but also introduces significant challenges for change detection (CD). For example, HRRS images, due to limited spectral information [6] or drastic spectral changes between changed and unchanged areas [7], present "intra-class variance" and "inter-class similarity", which negatively impacts the results of CD, leading to false alarms and missed detections of large-scale targets, as shown in Fig. 1.

For a long time, identifying changes in the same location across different periods has been a major research focus [8], [9]. Traditional CD methods rely on manually designed feature engineering, including Change Vector Analysis [10], Principal Component Analysis [11], etc., which have shown poor robustness in modern CD [12]–[14]. Many researchers have employed deep learning, including Convolutional Neural Networks (CNN) [15]–[17], Long Short-Term Memory networks [18]–[20], and Transformers [21]–[23], significantly advancing the field of RSCD. However, researchers have found that CNN are not adept at capturing long-range contextual information and less salient targets. Many efforts have been made to mitigate these issues and improve CD results. For instance, FRM-DeepLab [16] uses annotated samples to update a mask-based framework, designing an autoencoder to extract deep features from unmarked areas to provide more auxiliary information for MaskNet and reduce overfitting. L-UNet [20] incorporates gating mechanisms to replace convolutional layers, creating a convolutional LSTM structure that is added atop each encoding level. This design captures temporal relationships across different resolution levels. Although they have achieved good results, they still cannot fully explore the long-range contextual information in HRRS images, which is vital for good performance in RSCD. As interest in Transformers for RSCD continues to grow, ample experimental results have demonstrated their effectiveness in exploring global contextual information in HRRS image CD tasks [24], [25]. Despite showing great potential, they also bring huge computational costs and insufficient attention to local details [26], [27].

Additionally, although many current CD methods have shown good performance, most methods adopt complex and redundant structures for feature extraction, information interaction, or multi-scale fusion between bitemporal images [28], [29]. This greatly increases the complexity and scale of the models, hindering the application of CD models in industrial scenarios. In an era where massive remote sensing data is generated at every moment, large networks struggle to quickly and efficiently process this vast amount of data while preserving both high accuracy and fast computational speed. In Fig. 2, we show the F1 scores, model parameters (Params), and computational speed (Frames Per Second, FPS) of recently proposed RSCD methods, including the proposed method, SNUNet [30], TFI-GR [31], SEIFNet [32], SFEARNet [33], USSFCNet [28], TinyCD [29], BiT [34] and A2Net [35], evaluated on three RSCD datasets: HRCUS-CD [36], WHU-CD [15], and LEVIR-CD+ [37]. Among them, TFI-GR achieved an F1 score of 0.7062 on HRCUS-CD but at a significant computational cost, with 27M parameters and 105 FPS. USSFCNet achieved an F1 score of 0.6354 on HRCUS-CD with only 1.52M parameters, but the computational speed was only 61 FPS. This highlights the urgent need to develop RSCD methods that have fewer model parameters, faster computational speed, and good accuracy.

Building on the above research, we introduce a efficient network named CHLF-Net, which features Centralized High-Low Rank Representation (CHLR) and Lightweight Deep-Shallow Fusion Attention (LDSA). The proposed method adopts a U-shape network with a twin backbone network as the main framework. Inspired by [29], low-level features already possess sufficient expressive power, thus a lightweight, shallow network is employed to extract multilevel temporal features. Thereout, we introduce CHLR, which includes High-Rank Represented Aggregation (HRRA) and Low-Rank Represented Comparison (LRRC). HRRA captures local change features of bitemporal images through MLP, complementing LRRC. The low-rank matrix in LRRC models the context of the bitemporal images respectively, forcing network to store rich contextual features in a more abstract way. The features output by HRRA and LRRC are concatenated and fed into a U-shaped decoder to supplement detailed features. This approach effectively captures changes between the bitemporal phases and distinguishes unchanged background information, thereby mitigating the poor detection performance caused by 'intra-class variance' and 'inter-class similarity' in HRRS images. Additionally, we designed LDSA to embed deep high-level semantic features into each layer of the decoder in the form of attention maps, achieving a finer recovery of change information.

The main contributions of this paper are outlined below:

1) A lightweight CD network, namely CHLF-Net, is proposed to address the challenge of balancing performance and speed for HRRS imagery. By using a shallow backbone network to maintain a lower parameter count, CHLF-Net compensates for the loss of change information caused by lightweighting. This is achieved through Centralized Low-High Rank Representation (CHLR) and Lightweight Deep-Shallow Attention (LDSA), which help the shallow network supplement contextual information and reduce the impact of

background noise.

2) Centralized Low-High Rank Representation (CHLR) was developed to efficiently address the issue of contextual information loss in shallow networks for change detection (CD) tasks. CHLR consists of two parallel components: High-Rank Representation Aggregation (HRRA) and Low-Rank Representation Comparison (LRRC).By using the low-rank matrix of LRRC to learn the advanced semantics of bitemporal images, combined with HRRA composed of convolutional layers, high-quality change object features are obtained.

3) Lightweight Deep-Shallow Attention (LDSA) is proposed to efficiently integrate deep and multi-level shallow information. Unlike other methods that directly fuse multiple or even all scale features together, we only fuse the deepest features with other single-scale features to ensure the decoder remains lightweight and efficient.

4) A set of comprehensive experiments on three RSCD datasets was conducted to validate the advantages of CHLF-Net, which outperformed baseline models including BiT, TinyCD, A2Net, and TFI-GR in terms of performance, and surpassed all baselines in speed except for FC-Diff and TinyCD. CHLF-Net achieved a strong balance between performance and speed, demonstrating excellent robustness for HRRS imagery.

The rest of this paper is organized as below. Section II provides a review of previous related work on deep learning methods, as well as context modeling and multi-level fusion challenges in RSCD. Section III details the our method. In Section IV, we show and analyze the experimental results of proposed method. Section V contains additional discussions. Lastly, Section VI concludes the paper with a summary of our works.

II. Related Work

A. Deep Learning for RSCD

Thanks to the strong representation abilities of deep learning, a multitude of techniques have been developed and have become the leading solutions. RSCD methods using deep learning are typically categorized into two primary forms: metric-based methods and classification-based methods. A most intuitive way to distinguish between the two methods is whether the network determines the change map.

Generally speaking, metric-based methods attempt to project bitemporal images into a high-dimensional feature space with consistent latent representations. The goal is to learn a distance metric that is most suitable for CD, thereby effectively distinguishing between change and non-change areas. To achieve this goal, triplet loss [38], coupling function [39], and contrast loss [4], [40], [41] are often used as CD metric learning loss functions. DASNet [13] designed a convolution-based dual attention module acorss both spatial and channel dimensions, coupled with a weighted double-margin contrastive loss to emphasize change features, thereby alleviating data imbalance issues.

Classification-based methods, on the other hand, fuse bitemporal image features in some manner and directly classify the feature map into changed and unchanged categories through the network head, generating the final change map. In recent years, due to the progress of classification-based methods in other fields of computer vision, classification-based CD methods have become a research hotspot [42]–[48]. Caye Daudt et al. [42] constructed three classical CD models based on U-shape structure, which fuses bitemporal features through simple concatenation or difference operation. DTCDSCN [43] constructed a building CD and semantic segmentation model, obtaining object-level features through the semantic segmentation sub-model, thereby making the main task of CD more discriminative. FDCNN [44] used CNN to learn deep features through scene classification tasks in RS images with different spatial resolutions and applied them to a feature fusion network (FF-Net) through transfer learning, aiming to acquire more generalized change information.

Unlike the aforementioned CD methods, the method proposed in this paper aims to efficiently learn and utilize contextual dependencies in bitemporal images under limited computational cost. Consequently, our method demonstrates both rapid detection and high accuracy.

B. Context Modeling for RSCD

Context modeling is crucial for RSCD tasks, as it determines the ability to accurately segment the target changes of interest under complex backgrounds and image conditions. To effectively model contextual information, many studies concentrate on expanding the receptive field (RF) of networks, such as using deeper CNN models [38], [39], [44], employing atrous convolution [28], applying Transformers [21]–[23], [45], [46], and improving attention mechanisms [4], [47], [48]. For example, Zhang et al. [38] adopted the deep backbone network ResNet-101 [49] and used atrous convolution to produce denser feature maps. Luo et al. [50] found that in CNNs, An effective RF exhibits a Gaussian distribution and encompasses only a portion of the full theoretical receptive field. Therefore, more recent work focuses on applying Transformers and improving attention mechanisms. Wang et al. [46] developed a large-scale RS dataset and pre-trained ViTAEv2 [51], achieving excellent performance in CD tasks. Zhang et al. [48] designed an novel attention mechanism to fusion multi-level features, enhancing the internal integrity of objects in the change map.

Most previous methods tend to develop intricate network architectures to extract bitemporal features to capture contextual change information as much as possible. Unlike these RSCD methods, we designed a less parameter-intensive and computationally efficient CHLR to effectively learn and utilize contextual dependencies in bitemporal images.

C. Multi-level Feature Fusion for RSCD

Feature Fusion is a fundamental process for binary segmentation in CD tasks [31], [35], [52]. In convolutional backbone network methods, to avoid the loss of small targets and detailed information due to multiple downsampling operations, designing reasonable multi-level feature fusion becomes particularly important. For example, Li et al. [31] enhanced the

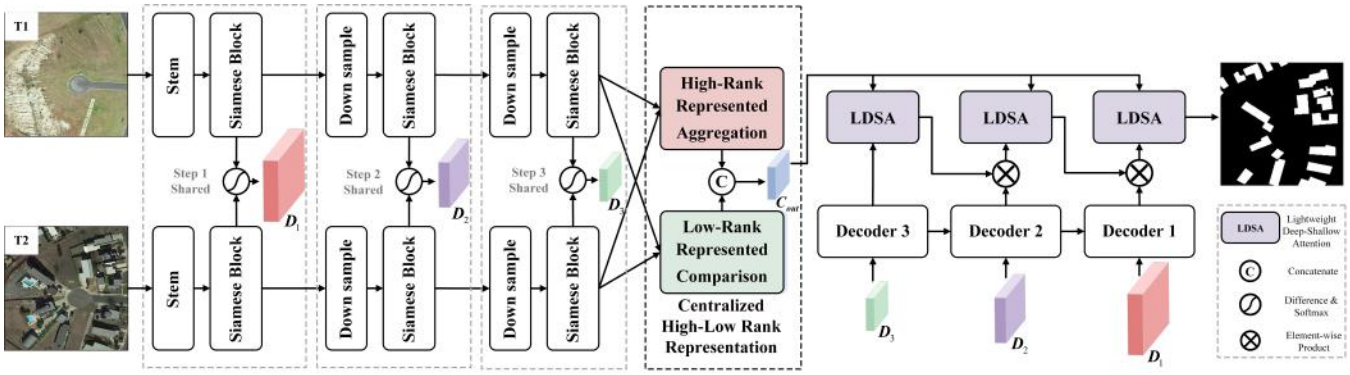


Fig. 3. Diagram of the CHLF-Net structure.

interaction between bitemporal features and performed multiple aggregations. Peng et al. [52] developed a dense attention mechanism comprising multiple upsampling attention units, aiming for high-level features to better guide low-level ones.

Unlike above CD methods, to reduce redundancy and computational cost, the proposed method fuses deep high-level features with multi-level features separately and transmits them upwards in the form of attention maps, efficiently generating the change map.

III. Methodology

In this section, we begin by examining the outright framework of the CHLF-Net and further explain the implementation details of the Centralized High-Low Rank Representation and Lightweight Deep-Shallow Fusion Attention.

A. Motivation and Framework Overview

RSCD identifies change areas from multitemporal remote sensing images, requiring attention to both rich visual and temporal information. In the same region across different temporal HRRS images, regardless of actual changes, there may be varying degrees of spectral differences, high complexity of the detected objects, and positional shifts due to misalignment. Therefore, a rigorous exploration of contextual information across and within temporal instances is crucial for high-quality RSCD.

The overall pipeline of our proposed method utilizes a U-Shape network, as shown in Fig. 3. Below, this article will introduce the encoder and decoder components individually.

1) Encoder: Our Siamese backbone utilizes the lightweight EfficientNetV2-S [53] network to extract multi-layer bitemporal features. Inspired by [29], to further reduce model parameters, we only used the first three stages of EfficientNetV2-S. For a registered bitemporal image pair T^1 and T^2, the extracted multilevel bitemporal features across three stages are represented as F_i^1 and F_i^2, $i \in \{1, 2, 3\}$.

To generate temporal difference features, we apply a feature difference method that involves an element-wise subtraction followed by a sigmoid activation function D_1, D_2, D_3 from bitemporal features as:

$$\bar{D}_i = \delta(F_i^1 \ominus F_i^2),\ i \in \{1, 2, 3\} \tag{1}$$

where $\delta(\cdot)$ and $\ominus$ denote the sigmoid activation function and element-wise subtraction operation, respectively.

To address the limited feature extraction capability of shallow backbone networks, we introduce CHLR to improve the network's representation power for HRRS and enhance its ability to discriminate temporal difference features. The bitemporal features F_3^1 and F_3^2 output by the backbone network Stage3 will serve as the input to CHLR. This procedure can be formulated as:

$$C_{out} = \mathcal{C}(F_3^1,\ F_3^2) \tag{2}$$

where $\mathcal{C}(\cdot)$ represents the function of CHLR.

2) Decoder: To maintain balance with the encoder, we designed a lightweight decoder that fuses multilevel features in a down-top manner, gradually recovering the fine details of the predicted change map. We stack convolutional layers and upsampling layers to gradually restore the size of the difference feature map. Furthermore, LDSA is designed to efficiently fuse deep and other-level feature maps. The particular of LDSA are presented in Section III-C. The upsampling process can be expressed as follows:

$$\begin{cases} \bar{\bar{D}}_3 = \mathrm{Conv}_{1\times1}\{\mathrm{Up}(C_{out})\} \\ \bar{\bar{D}}_{i-1} = \mathrm{Conv}_{1\times1}\{\mathrm{Up}(D_i)\},\ i = 3, 2, 1 \end{cases} \tag{3}$$

where $Conv_{1\times1}(\cdot)$ denotes a 1×1 convolutional layer with LeakyReLU activation function and batch normalization. $Up(\cdot)$ represents that the difference map is upsampled to twice its original resolution through bilinear interpolation. C_{out} denotes the output features of CHLR. D_i denotes the tensor obtained by skip connection of two levels of features. Subsequently, the upsampled features are skip-connected with the temporal difference features of the previous level, formulated as follows:

$$D_i = \mathrm{Conv}_{3\times3}(\bar{D}_i \oplus \bar{\bar{D}}_i),\ i = 3, 2, 1 \tag{4}$$

where $\oplus$ denotes the element-wise addition operation. Subsequently, LDSA fuses the deepest features with other shallow features to obtain an attention mask, which is then multiplied and propagated upwards step by step. This can be formulated as follows:

$$\begin{cases} \hat{D}_2 = \mathcal{D}(D_3, C_{out}) \\ \hat{D}_{i-1} = D_{i-1} \otimes \mathcal{D}(D_i, C_{out}),\ i = 2, 1 \end{cases} \tag{5}$$

where $Conv_{3\times3}(\cdot)$ signifies a 3×3 convolutional layer combined with LeakyReLU activation function and batch nor-

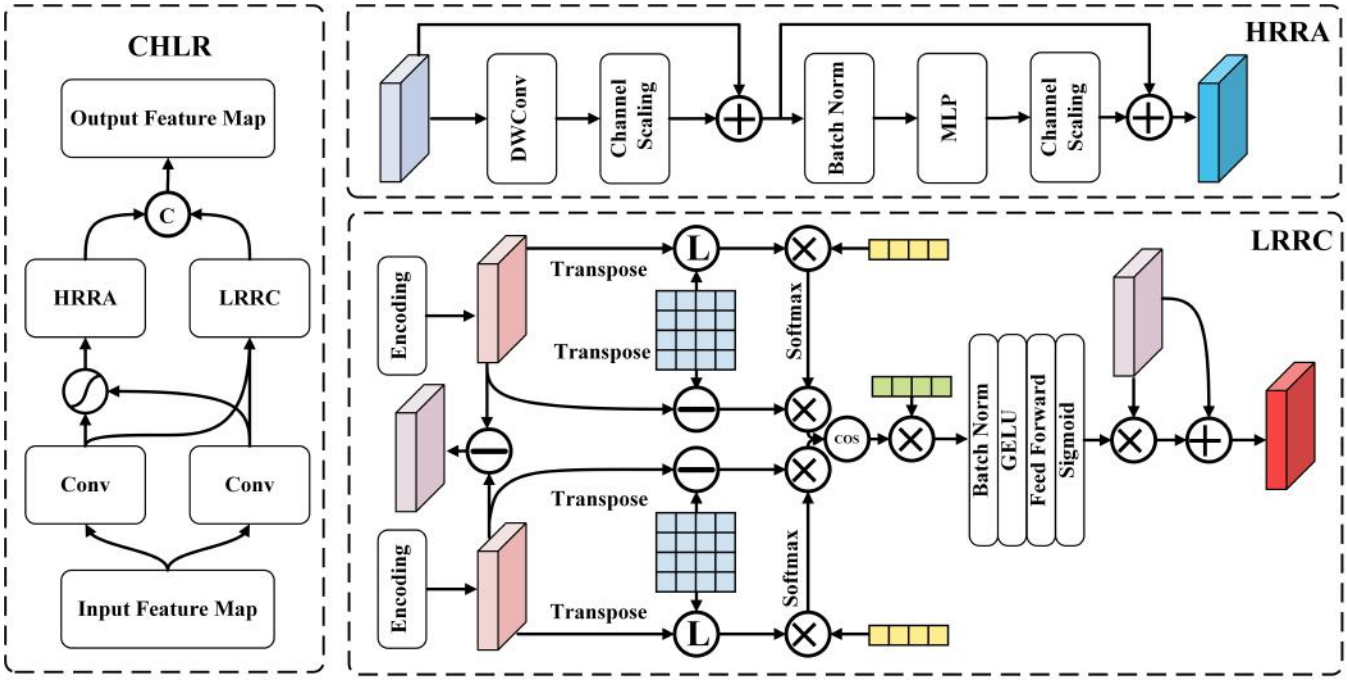


Fig. 4. Diagram of the CHLR structure.

malization, $\otimes$ denotes element-wise multiplication, and $\mathcal{D}(\cdot)$ represents the function of LDSA. In such a way, the low-level detailed features from the shallow network is efficiently and lightly fused with the rich high-level semantic information, gradually generating the final change map.

B. Centralized High-Low Rank Representation

As illustrated in Fig. 4, our proposed CHLR has two parallel branches, namely HRRA and LRRC. HRRA captures long-range dependencies of cross-temporal features (i.e., global information). Meanwhile, to extract high-quality single-temporal semantic information, we propose LRRC, which aggregates high-rank bitemporal features using two learnable low-rank parameter matrices, allowing the model to learn more critical semantic information. The defference maps of these two modules are concatenated along the channels dimension to serve as the output of CHLR, which is then used for the subsequent decoding of the change feature map. Before and after entering HRRA and LRRC, we use large kernel convolution blocks as Stem blocks to smooth the bitemporal features respectively, formulated as follows:

$$C_{in}^{i} = \mathcal{S}(F_3^i),\ i \in \{1, 2\} \tag{6}$$

$$C_{out} = \mathrm{Cat}(\mathcal{H}(C_{in}^1, C_{in}^2), \mathcal{L}(C_{in}^1, C_{in}^2)) \tag{7}$$

where $\mathcal{S}(\cdot)$ denotes the Stem block, which includes of a 7×7 convolution with 64 output channel, succeeded by GELU activation function and Batch Normalization. $\mathcal{H}(\cdot)$ represents HRRA, and $\mathcal{L}(\cdot)$ represents LRRC.

1) High-Rank Represented Aggregation (HRRA): Recently, vision MLP networks have received widespread attention. These networks follow the traditional ViT design architecture and use multi-layer perceptrons (MLP) to replace complex attention mechanisms, modeling global spatial information to capture long-range dependencies. The proposed HRRA employs 1×1 depthwise convolution and MLP to ensure the global modeling capability of the high-rank module. Between these two modules, batch normalization and channel scaling operations are inserted to enhance the model's generalization and robustness. Specifically, the outputs C_{in}^1 and C_{in}^2 from the Stem are processed through element-wise subtraction and sigmoid activation functions, and then fed into depthwise convolution. In contrast to standard convolution, depthwise convolution considerably decreases the parameters while enhancing feature representation capability. Additionally, to boost generalization, channel scaling and a residual connection are included. This process can be formulated as follows:

$$C_{in}^h = C_{in}^1 \ominus C_{in}^2 \tag{8}$$

$$\hat{C}_{in} = \mathrm{CS}(\mathrm{DConv}(C_{in}^h)) + C_{in}^h \tag{9}$$

where $\mathrm{CS}(\cdot)$ is the Channel Scaling. $\mathrm{DConv}(\cdot)$ denotes depthwise convolution with a 1×1 kernel size.

Before entering the MLP, the feature output $\hat{C}_{in}$ from the depthwise convolution is first fed into a batch normalization layer and then into an MLP consisting of two 1×1 convolutional layers with GELU activation functions in between. This method reduces computational cost while enhancing feature representation capability, satisfying the demands of vision tasks. Subsequently, channel scaling and a skip connection are similarly introduced sequentially. This process can be formulated as follows:

$$\mathcal{H}(C_{in}^1, C_{in}^2) = \mathrm{CS}(\mathrm{BN}(\mathrm{MLP}(\hat{C}_{in}))) + \hat{C}_{in} \tag{10}$$

2) Low-Rank Represented Comparison (LRRC): LRRC first maps $C_{in}^1, C_{in}^2 \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ to $\widehat{C}_{in}^1 = \{\hat{c}_1^1, \hat{c}_2^1, \cdots, \hat{c}_N^1\}$ and $\widehat{C}_{in}^2 = \{\hat{c}_1^2, \hat{c}_2^2, \cdots, \hat{c}_N^2\}$, where $\mathrm{N} = \mathrm{H} \times \mathrm{W}$ is the number of bitemporal spatial features. Then, $\widehat{C}_{in}^1$ and $\widehat{C}_{in}^2$ are

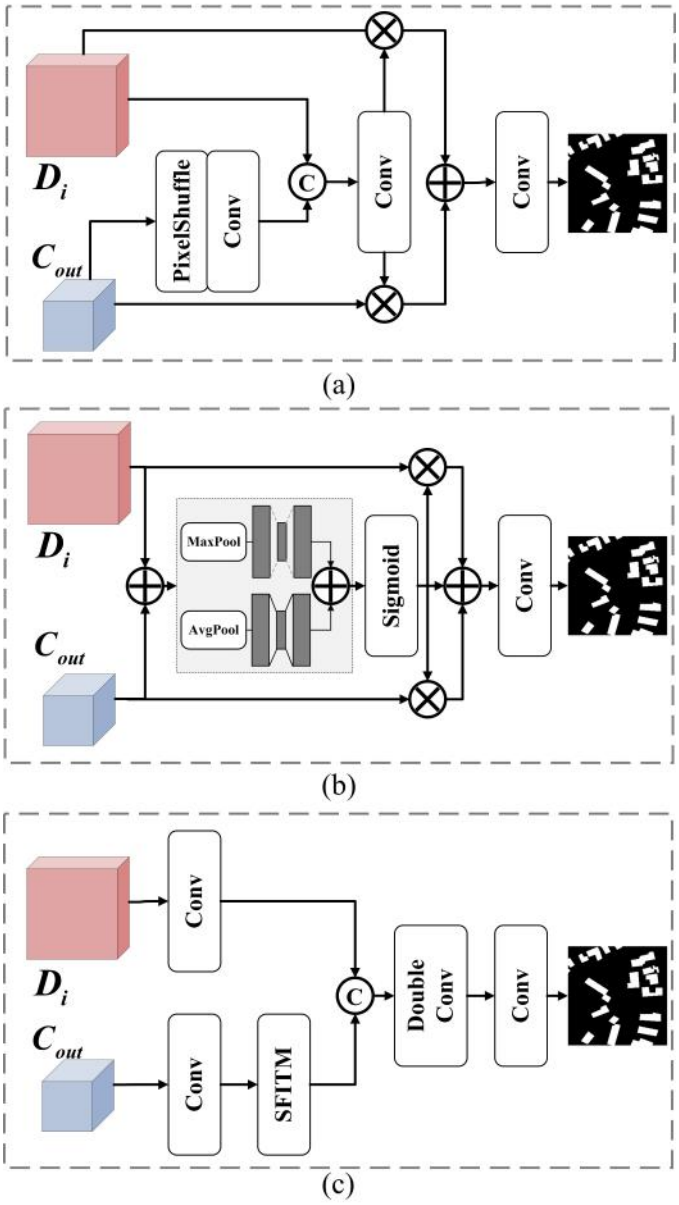


Fig. 5. Architectural comparison of the different feature fusion modules. (a) The proposed LDSA. (b) Adapted ACFM with channel-spatial attention. (c) Adapted EARM, which utilizes its original SFITM and double convolution components.

mapped to general variation features C_{in}^l through a general difference operation for subsequent processing. Next, $\widehat{C}_{in}^1$ and $\widehat{C}_{in}^2$ respectively learn two sets of important parameters and one shared parameter for the corresponding bitemporal features: (1) low-rank matrices $L^1 = \left\{l_1^1, l_2^1, \cdots, l_K^1\right\}$ and $L^2 = \left\{l_1^2, l_2^2, \cdots, l_K^2\right\}$, where K is the intrinsic dimensionality factor. (2) Smoothing parameters $s^1 = \left\{s_1^1, s_2^1, \cdots, s_K^1\right\}$ and $s^2 = \left\{s_1^2, s_2^2, \cdots, s_K^2\right\}$. (3) Shared smoothing parameters $s^3 = \{s_1, s_2, \cdots, s_C\}$.

Next, we use C_{in}^1 as an example for detailed explanation. C_{in}^1 is encoded through a set of residual convolutional groups. The main branch uses 1×1 convolutions, 3×3 convolutions, and 1×1 convolutions, while the shortcut branch uses 1×1 convolutions. The tensors obtained from the main branch and the shortcut branch are then added together to obtain $\widehat{C}_{in}^1$. The first 1×1 convolution in the main branch is used for dimensionality reduction, and the second 1×1 convolution is used for dimensionality restoration. Each convolution layer includes Batch Normalization and GELU activation functions. This residual convolutional group does not change the feature dimensions but helps the model to converge better with a smaller computational cost.

Subsequently, the low-rank matrix $L^1 = \left\{l_1^1, l_2^1, \cdots, l_K^1\right\}$ which is learnable is matched with $\widehat{C}_{in}^1 = \left\{\hat{c}_1^1, \hat{c}_2^1, \cdots, \hat{c}_N^1\right\}$ to perform a difference operation, with the smoothing parameters $s^1 = \left\{s_1^1, s_2^1, \cdots, s_K^1\right\}$ assisting the model in better convergence. This allows us to generate learnable weights from the difference $\hat{C}_n^1 - l_k$ between the low-rank matrix and the feature tensor, and after multiplying the difference by these weights and summing along the dimension N, we obtain $E^1 = \left\{e_1^1, e_2^1, \cdots, e_K^1\right\}$. This process forces the bitemporal features to refine the most important information through the low-rank matrix. The quantity e_k^1 in the process described above can be formalized as follows:

$$q_{nk} = e^{-s_k \left\| \hat{C}_n^1 - l_k \right\|} \tag{11}$$

$$p_n = \sum_{j=1}^{K} e^{-s_j \left\| \hat{C}_n^1 - l_j \right\|} \tag{12}$$

$$e_k^1 = \sum_{i=1}^{N} \frac{q_{ik}}{p_i} \left(\hat{C}_i^1 - l_k \right)^2 \tag{13}$$

Among them, $k \in \{1, 2, \cdots, K\}$, $n \in \{1, 2, \cdots, N\}$, and $\|\cdot\|$ represents the calculation of the L2 norm. Through the above steps, we obtain $E^1 = \{e_1^1, e_2^1, \cdots, e_K^1\}$, where $e_k^1 = \mathbb{R}^{1 \times C}$, which contains information about the entire image T^1. Meanwhile, $\widehat{C}_{in}^2$ uses L^2 and s^2 to obtain $E^2 = \{e_1^2, e_2^2, \cdots, e_K^2\}$ through the same steps, where $e_k^2 = \mathbb{R}^{1 \times C}$, which contains information about the entire image T^2. Afterwards, we will fuse e^1 and e^2 across the K scalar dimensions to obtain the highly abstract change information E, which is a C-dimensional vector, and can be formulated as:

$$E = s^3 \cos(E^1, E^2) \tag{14}$$

where s^3 is a shared smoothing parameter, which is multiplied by the result after the fusion of bitemporal features. $\cos(\cdot)$ denotes the cosine similarity operation performed across the K dimensions to fuse the bitemporal features. At this point, E captures the most important change features. However, to recover the change map to its original size while retaining fine details, E needs to be fused with the original input features. After normalizing E, we apply a linear layer to transform the map E into dimensions of C$\times$1$\times$1. Subsequently, an activation function is applied, and a multiplication is performed with the general change features C_{in}^l along the channel dimension. The resulting features are then added back to C_{in}^l to prevent degradation during training. The formula is as follows:

$$Z = C_{in}^l \otimes (\delta(\mathrm{FC}(E))) \tag{15}$$

$$\mathcal{L}(C_{in}^1, C_{in}^2) = C_{in}^l \oplus Z \tag{16}$$

where $\mathrm{FC}(\cdot)$ represents a linear layer that includes a batch normalization, a GELU activation function, and a Linear layer. $\otimes$

denotes element-wise multiplication with broadcasting along the channel dimension, $\delta(\cdot)$ indicates the sigmoid activation function, and $\oplus$ indicates element-wise addition.

C. Lightweight Deep-Shallow Attention

Several studies suggest that multi-scale features contain different information about the object of interest (e.g., deep network features include semantic information, whereas shallow network features contain detail information). In addition to the skip connections in the U-Shape network, we have incorporated LDSA. During the process of fusing multi-scale features, for minimize computational cost, we prefer to fuse only the deepest features with features from other steps, thereby enhancing the performance on change features. The proposed LDSA is illustrated in Fig. 5 (a).

LDSA is used in the decoder stage to fuse the output C_{out} of CHLR with D_i, where $i \in \{1, 2, 3\}$. Unlike most other works that use linear interpolation methods, our upsampling operation employs Pixel Shuffle. First, Pixel Shuffle is used to upsample C_{out} to the same resolution as the D_i to be fused, while a 3×3 convolutional layer is applied to expand the channels, resulting in the upsampled feature D_4. Subsequently, D_i is concatenated with D_4 to obtain the deep-to-shallow coarse feature D. We then perform element-wise multiplication between D and D_i, D_4, followed by addition, and apply a 1×1 convolutional to transform the features to a pixel attention mask M_i with a shape of $1 \times \mathrm{H} \times \mathrm{W}$. The proposed LDSA can be formulated as follows:

$$D_4 = \mathrm{Conv}_{3\times3}(\mathrm{PS}(C_{out})) \tag{17}$$

$$D = \mathrm{Conv}_{3\times3}(\mathrm{Cat}(D_i, D_4)) \tag{18}$$

$$\mathcal{D}(D_i, C_{out}) = \mathrm{Conv}_{1\times1}((D \otimes D_i) \oplus (D \otimes D_4)) \tag{19}$$

Where $\mathrm{PS}(\cdot)$ represents the Pixel Shuffle operation, $\mathrm{Cat}(\cdot)$ represents the concatenation operation. Ultimately, three attention masks are obtained, and as illustrated in Fig. 3, the semantics obtained from the deepest layer are progressively propagated upwards.

D. Deep Supervised Learning

Regarding the severe data imbalance issue in CD [30], [35], we employ a hybrid loss function composed of the BCE loss $\mathrm{L_{BCE}}$ and the DICE [54] loss $\mathrm{L_{DICE}}$.

The hybrid loss function can be formulated as:

$$\mathrm{L}(Y, \hat{Y}) = \sum_{j=1}^{3} \left(\mathrm{L_{BCE}}\left(Y, \hat{Y}_i\right) + \mathrm{L_{DICE}}\left(Y, \hat{Y}_i\right) \right) \tag{20}$$

where $\hat{Y}_j,\ j \in \{1, 2, 3\}$, denotes the predicted change maps for three levels, Y and $\hat{Y}$ represent the label and the corresponding predicted change map, respectively.

IV. Experiments

A. Experimental Setup

1) Implementation Details: Our method uses EfficientNetv2-S [53] with ImageNet pre-trained weights as the backbone. We use the Adam optimizer with a momentum of 9e-1 and a weight decay to 1e-4. The initial learning rate was set to 5e-4, using a Poly learning strategy to expedite network convergence. The batch size was 32 with 40,000 iterations. We employ three different techniques for data augmentation, including random flipping, cropping, and temporal iamge exchanging. Both training and inference are performed on a single Nvidia GeForce RTX 4090.

2) Quantitative Evaluation: We assess performance of proposed model using standard CD metrics, including precision (Pr), recall (Re), intersection over union (IoU), F1 score ($F1$) and Kappa coefficient ($Kappa$), formulated as follows:

$$\mathrm{Pr} = \frac{Tp}{Tp + Fp} \tag{21}$$

$$\mathrm{Re} = \frac{Tp}{Tp + Fn} \tag{22}$$

$$F1 = 2\frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{23}$$

$$IoU = \frac{Tp}{Tp + Fp} \tag{24}$$

$$OA = \frac{Tp + Tn}{Tp + Tn + Fn + Fp} \tag{25}$$

$$P_e = \frac{(Tn + Fn) \cdot (Tn + Fp) + (Fp + Tp) \cdot (Fn + Tp)}{(Tp + Fp + Tn + Fn)^2} \tag{26}$$

$$Kappa = \frac{OA - P_e}{1 - P_e} \tag{27}$$

Where Tp, Tn, Fp, and Fn represent the predicted pixels of true positive, true negative, false positive, and false negative, respectively. Additionally, model efficiency is further quantified using Latency, Frames Per Second (FPS), FLOPs, and parameters (Param.).

B. Datasets

1) HRCUS-CD [36]: This dataset comprises 11,388 pairs of HRRS images, including over 12,000 annotated instances. These images are from Zhuhai, China, an area undergoing urbanization and industrialization, and they include urban built-up areas, farmland, and mountainous regions, providing good diversity. The dataset is split into 7:2:1 ratios, with 7,974/2,276/1,138 image pairs used for training/validation/testing, respectively.

2) LEVIR-CD+ [37]: This dataset is an extended version of LEVIR-CD [4], comprising a total of 985 pairs of 1024×1024 HRRS images, which are primarily from 20 different areas in Texas, USA, with a greater focus on urban areas under seasonal and lighting changes. To ensure consistency with previous experiments, all images are divided into non-overlapping 256×256 patches, resulting in 10,192/5,568 image pairs for training/testing.

3) WHU-CD [15]: This dataset comprises a pair of HRRS images with a size of $32,507 \times 15,354$. It features diverse building types with varying colors, sizes, and functions, along with numerous large vehicles, containers, and artificial non-building structures such as greenhouses. It is an ideal dataset for assessing the potential of CD models. Due to limited

TABLE I
COMPARISON WITH THE NINE DIFFERENT METHODS IN TERMS OF PRE, REC, F1, IOU, KAPPA ON THREE DATASETS. ALL THE SCORES ARE DESCRIBED IN PERCENTAGE (%). COLOR CONVENTION: BEST, 2ND-BEST.

Method	HRCUS-CD					WHU-CD					LEVIR-CD+				
	Pr.	Re.	F1	IoU	Kappa	Pr.	Re.	F1	IoU	Kappa	Pr.	Re.	F1	IoU	Kappa
FC-Diff	0.4724	0.4978	0.4848	0.3199	0.4747	0.8712	0.7243	0.7910	0.6543	0.7822	0.7759	0.7795	0.7777	0.6363	0.7682
SNUNet	0.7652	0.5585	0.6457	0.4768	0.6400	0.9536	0.8952	0.9235	0.8578	0.9200	0.7835	0.7895	0.7865	0.6481	0.7774
TFI-GR	0.7062	0.7416	0.7235	0.5668	0.7181	0.9255	0.9232	0.9243	0.8593	0.9208	0.7482	0.8166	0.7809	0.6405	0.7712
SEIFNet	0.7724	0.6507	0.7064	0.5460	0.7012	0.9046	0.8863	0.8953	0.8105	0.8905	0.8550	0.7850	0.8145	0.6927	0.8111
SFEARNet	0.7546	0.7222	0.7380	0.5848	0.7332	0.9364	0.9157	0.9259	0.8621	0.9225	0.8319	0.8117	0.8217	0.6973	0.8142
USSFCNet	0.6354	0.7149	0.6728	0.5069	0.6662	0.9246	0.8133	0.8654	0.7627	0.8595	0.8273	0.7315	0.7765	0.6346	0.7676
TinyCD	0.6860	0.7010	0.6934	0.5307	0.6875	0.9313	0.8497	0.8886	0.7995	0.8837	0.7942	0.7925	0.7933	0.6575	0.7846
BiT	0.7049	0.7332	0.7188	0.5610	0.7133	0.9546	0.9010	0.9270	0.8640	0.9238	0.8295	0.7650	0.7960	0.6611	0.7877
A2Net	0.7611	0.7378	0.7492	0.5990	0.7445	0.9454	0.9191	0.9321	0.8728	0.9290	0.8232	0.8146	0.8189	0.6933	0.8112
Ours	0.7843	0.7309	0.7567	0.6086	0.7522	0.9651	0.9138	0.9388	0.8846	0.9360	0.8154	0.8339	0.8245	0.7014	0.8170

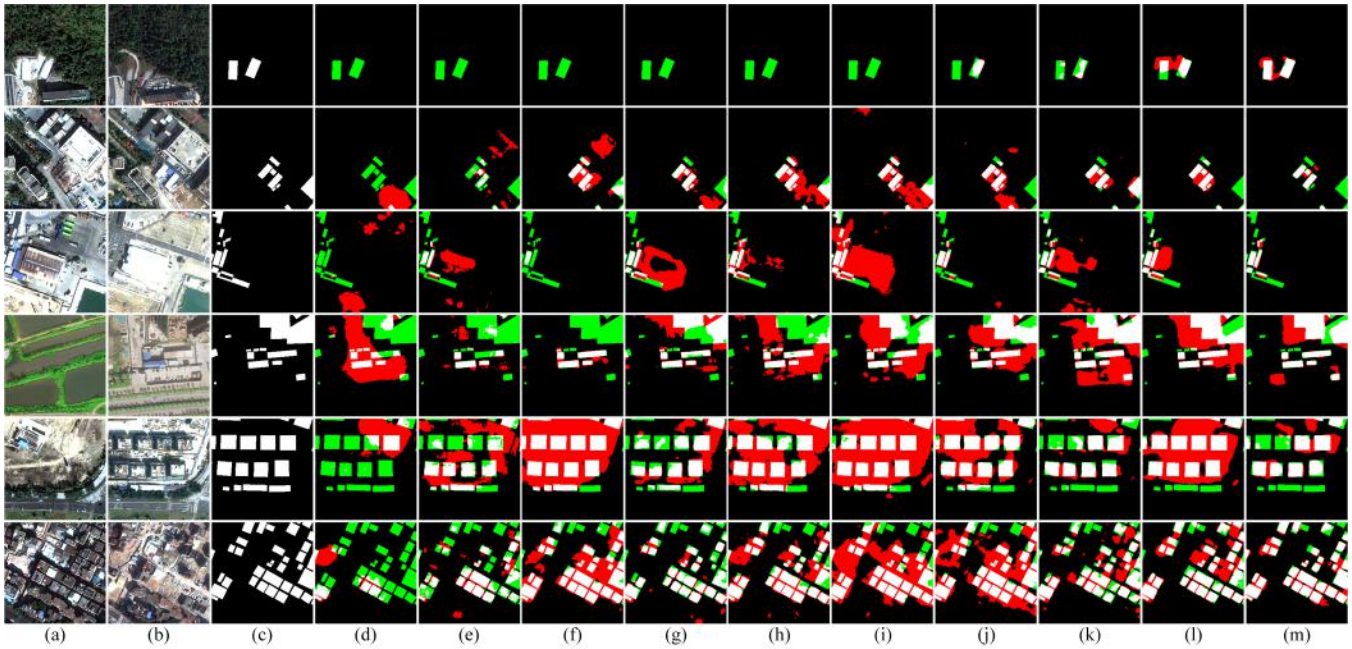


Fig. 6. Visual comparisons of the proposed method and the different methods on the HRCUS-CD dataset. (a) t1 images. (b) t2 images. (c) Ground truth. (d) FC-Diff. (e) SNUNet. (f) TFI-GR. (g)SEIFNet. (h) SFEARNet. (i)USSFCNet. (j) TinyCD. (k) BiT. (l) A2Net. (m) Ours. The rendered colors represent true positives (white), false positives (red), true negatives (black), and false negatives (green).

TABLE II
COMPARISON WITH THE NINE METHODS IN TERMS OF INFERENCE TIME, FPS, FLOPS, PAREM.

Method	Inference Time(ms)	FPS	FLOPs(G)	Parem.(M)
FC-Diff	0.56	1768.22	5.32	1.55
SNUNet	7.66	130.45	54.82	12.03
TFI-GR	9.50	105.16	9.73	27.78
SEIFNet	17.20	58.14	8.37	27.91
SFEARNet	43.22	23.14	4.65	5.56
USSFCNet	16.25	61.53	4.86	1.52
TinyCD	3.63	245.09	1.54	0.29
BiT	5.37	186.04	16.88	3.01
A2Net	7.26	137.68	6.02	3.78
Ours	4.49	222.87	4.96	1.27

analysis of the dataset splits in [15], following other works [22], [26], [34], [35], we partitioned the original image are divided into non-overlapping 256×256 patches and randomly split them into 6,095/762/763 image pairs for training/validation/testing, respectively.

C. Comparison with Various Methods

In this work, we assess our model againt Nine different RSCD methods, including two relatively large models, SNUNet, TFI-GR, SEIFNet and SFEARNet, and five smaller models, FC-Diff, USSFCNet, TinyCD, BiT, and A2Net. We replicated their publicly available code and aligned their hyperparameters with the experimental settings of this work. We also provide a report on the computational costs and parameters specific to the same machine.

1) Quantitative Analysis:

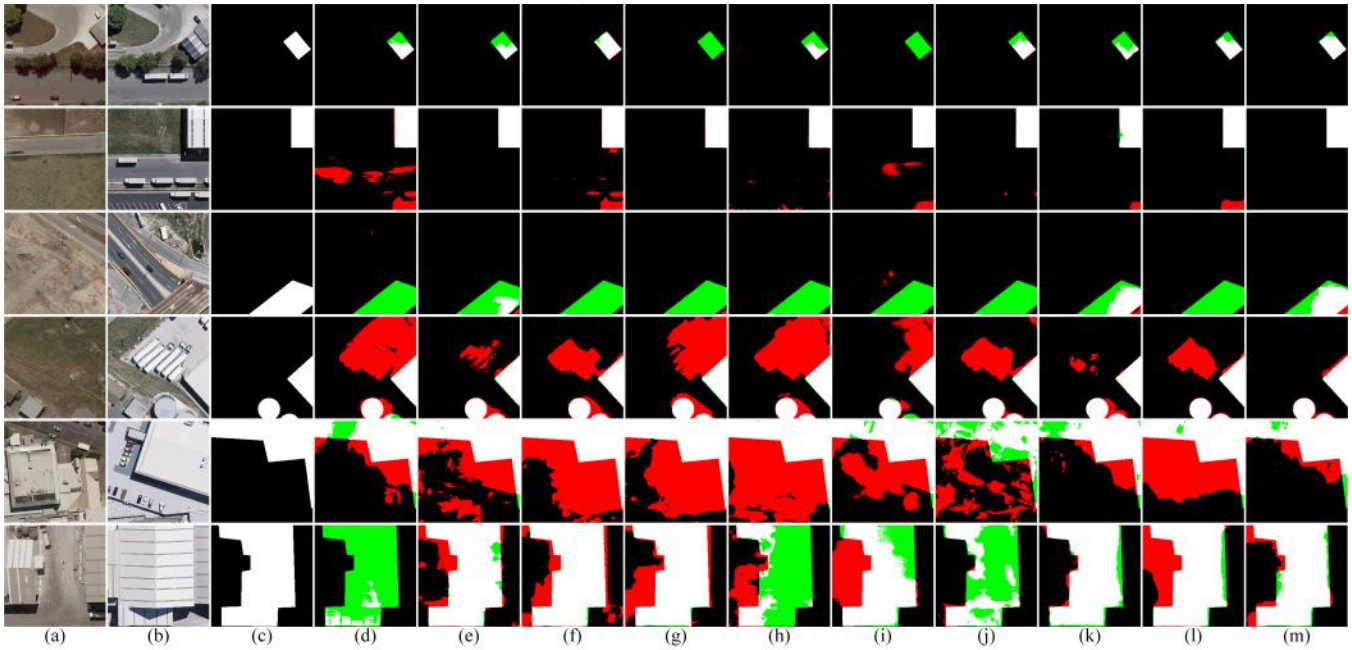


Fig. 7. Visual comparisons of the proposed method and the different methods on the WHU-CD dataset. (a) t1 images. (b) t2 images. (c) Ground truth. (d) FC-Diff. (e)SNUNet. (f) TFI-GR. (g)SEIFNet. (h) SFEARNet. (i)USSFCNet. (j) TinyCD. (k) BiT. (l) A2Net. (m) Ours. The rendered colors represent true positives (white), false positives (red), true negatives (black), and false negatives (green).

Table I provides a quantitative comparison of the our method against nine other methods across three RSCD datasets. The findings show that the our method attains the highest F1 score across all three datasets. Compared to the other five models, the proposed method shows an average improvement of 6.58% in F1 Score on the HRCUS-CD, 3.30% on the WHU-CD, and 3.78% on the LEVIR-CD+, except for the least-performing FC-Diff and the second-best A2Net.

Additionally, Table II reports the computational costs (Latency, FPS, FLOPs) and parameters (Param.) our method compared to others. The findings indicate that the our method requires less computational cost and has fewer parameters, showing only marginally poorer performance compared to TinyCD. Compared to other small parameter models like USSFCNet, which has a smaller parameters but poorer results in inference time and FPS, the proposed method shows more consistent performance.

On the HRCUS-CD, WHU-CD, and LEVIR-CD+ datasets, TFI-GR requires 27.78M parameters and 9.74G FLOPs to achieve F1 scores of 0.7235, 0.9243, and 0.7809, while the proposed method requires only 1.27M parameters and 4.96G FLOPs, achieving improvements of 3.32%, 1.45%, and 4.36% respectively. Additionally, TinyCD, with 0.29M parameters and 275FPS, achieves F1 scores of 0.6934, 0.8886, and 0.7933, while the proposed method, with a slight increase of 0.98M parameters and a decrease of 52FPS, achieves F1 scores of 0.7567, 0.9388, and 0.8245. These analyses highlight the Superiority of the our method as a lightweight model.

2) Qualitative Analysis: Fig. 6, 7 and 8 provide a visual analysis of the eight methods across the three datasets. In scenes with small targets, as seen the first row of Fig. 6 and the first row of Fig. 7, our model is more sensitive in detecting small-scale change targets compared to other models. In scenarios lacking contextual information, as seen the sixth rows of Fig. 7 and the fifth row of Fig. 7, the proposed model demonstrates good generalization, with smoother detection boundaries and more complete detection within the interiors of large buildings.

In the second and fourth rows of Fig. 7, numerous large vehicles present substantial interference for CD focused on buildings in the urban scenes of the WHU-CD dataset. The proposed model effectively distinguishes between the background and the target of interest, better separating building instances.

The second and third rows of Fig. 8 show numerous building change targets, densely packed, with varying degrees of misalignment. Other methods fail to exclude these interferences, leading to many false alarms. In contrast, our method accurately captures the semantics of the changing buildings and identifies the genuinely changed building areas more precisely.

In the HRCUS-CD dataset, although it also consists of urban scenes like the other two datasets, it exhibits significant spectral changes. This is one of the reasons why all methods generally show lower accuracy on this dataset. As shown in the third, fifth, and sixth rows of Fig. 6, it is evident that radiation significantly impacts, causing models like TFI-GR, USSFCNet, and TinyCD to struggle in distinguishing the accurate boundaries between the detection targets and irrelevant background, resulting in extensive false detections. Conversely, the proposed method can identify changing building targets more accurately, even under severe radiation interference.

A closer analysis reveals our model's strength in achieving a superior balance between recall (detecting true changes) and precision (avoiding false alarms). On the HRCUS-CD and WHU-CD, our model avoids the extensive false positives that affect methods with higher recall like TFI-GR, as visually evident in Fig. 6 and 7. In contrast, on the LEVIR-CD+, our model excels by attaining the highest recall, demonstrating

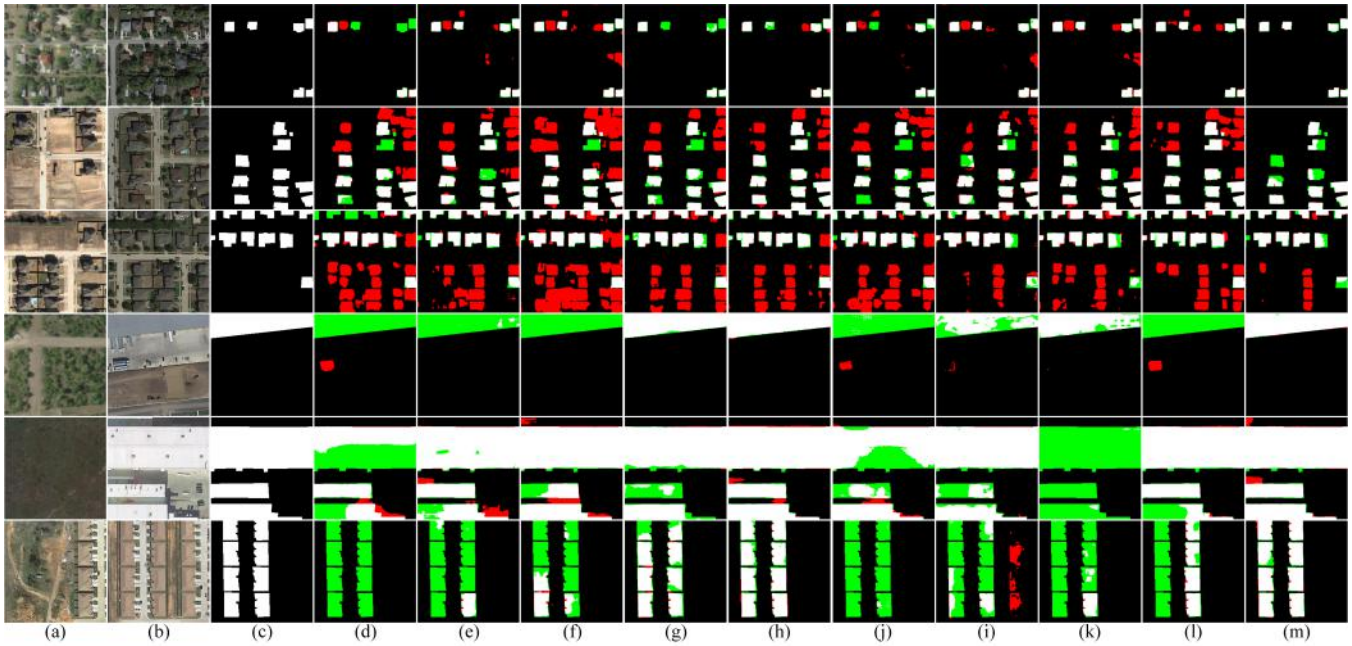


Fig. 8. Visual comparisons of the proposed method and the different methods on the LEVIR-CD+ dataset. (a) t1 images. (b) t2 images. (c) Ground truth. (d) FC-Diff. (e)SNUNet. (f) TFI-GR. (g)SEIFNet. (h) SFEARNet. (i)USSFCNet. (j) TinyCD. (k) BiT. (l) A2Net. (m) Ours. The rendered colors represent true positives (white), false positives (red), true negatives (black), and false negatives (green).

TABLE III
ABLATION STUDY ON THE PROPOSED METHOD SUPERIORITY IN TERMS OF F1, IOU, KAPPA ON THREE DATASETS. ALL THE SCORES ARE DESCRIBED IN PERCENTAGE (%). COLOR CONVENTION: BEST, 2ND-BEST.

No.	LDSA	LRRC	HRRC	HRCUS-CD			WHU-CD			LEVIR-CD+		
				F1	IoU	Kappa	F1	IoU	Kappa	F1	IoU	Kappa
#1	✘	✘	✘	0.6787	0.5137	0.6725	0.9091	0.8334	0.9050	0.7855	0.6467	0.7763
#2	✔	✘	✘	0.7277	0.5719	0.7227	0.9305	0.8700	0.9273	0.8052	0.6740	0.7968
#3	✘	✔	✘	0.7255	0.5693	0.7204	0.9227	0.8565	0.9192	0.8042	0.6725	0.7957
#4	✘	✘	✔	0.6922	0.5293	0.6863	0.9181	0.8486	0.9144	0.7948	0.6595	0.7862
#5	✘	✔	✔	0.7089	0.5403	0.6648	0.9196	0.8511	0.9160	0.7986	0.6647	0.7900
#6	✔	✘	✔	0.7371	0.5837	0.7324	0.9346	0.8773	0.9317	0.8082	0.6781	0.8001
#7	✔	✔	✘	0.7381	0.5849	0.7333	0.9330	0.8745	0.9300	0.8130	0.6849	0.8049
Ours	✔	✔	✔	0.7567	0.6086	0.7522	0.9388	0.8846	0.9360	0.8245	0.7014	0.8170

TABLE IV
EFFECT OF THE NUMBER OF INTRINSIC DIMENSIONALITY FACTOR K IN LRRC ON HRCUS-CD DATASETS. COLOR CONVENTION: BEST, 2ND-BEST.

K	Latency	FPS	F1	IoU	Kappa
16	4.54	220.33	0.7425	0.5905	0.7377
32	4.55	219.94	0.7521	0.6016	0.7467
64	4.49	222.87	0.7567	0.6086	0.7522
128	4.93	202.59	0.7525	0.6033	0.7480
256	5.37	186.22	0.7516	0.6021	0.7470

its capability to capture change features effectively without a significant drop in precision. This consistent ability to find an equilibrium is quantitatively confirmed by our model's top-ranking F1-score, IoU, and Kappa coefficients across all three datasets, as shown in Table I.

D. Ablation Studies

The ablation experiment in this study aims to validate the superiority of two important modules in the proposed method: CHLR and LDSA. Since CHLR also includes two components: HRRA and LRRC, and there is not much interaction between the two components, they are considered independent modules. We performed an extensive ablation study across three RSCD datasets.

1) Experiment Settings: To intuitively verify the effectiveness of each module, we bypassed CHLR for the features output by stage3, and passed them directly through LDSA to restore to the original size. This network serves as the basic BaseLine (#1 on Table III). Additionally, by isolating CHLR or LDSA through the aforementioned method, we obtained ablation BaseLines (#2 and #5 on Table III) for comparison with HRRA and CHLR. For the ablation setting of LRRC, we passed the feature tensor through the residual convolution block, then through LRRC without the cat operation, thus obtaining ablation BaseLines (#3 and #6 on Table III) for

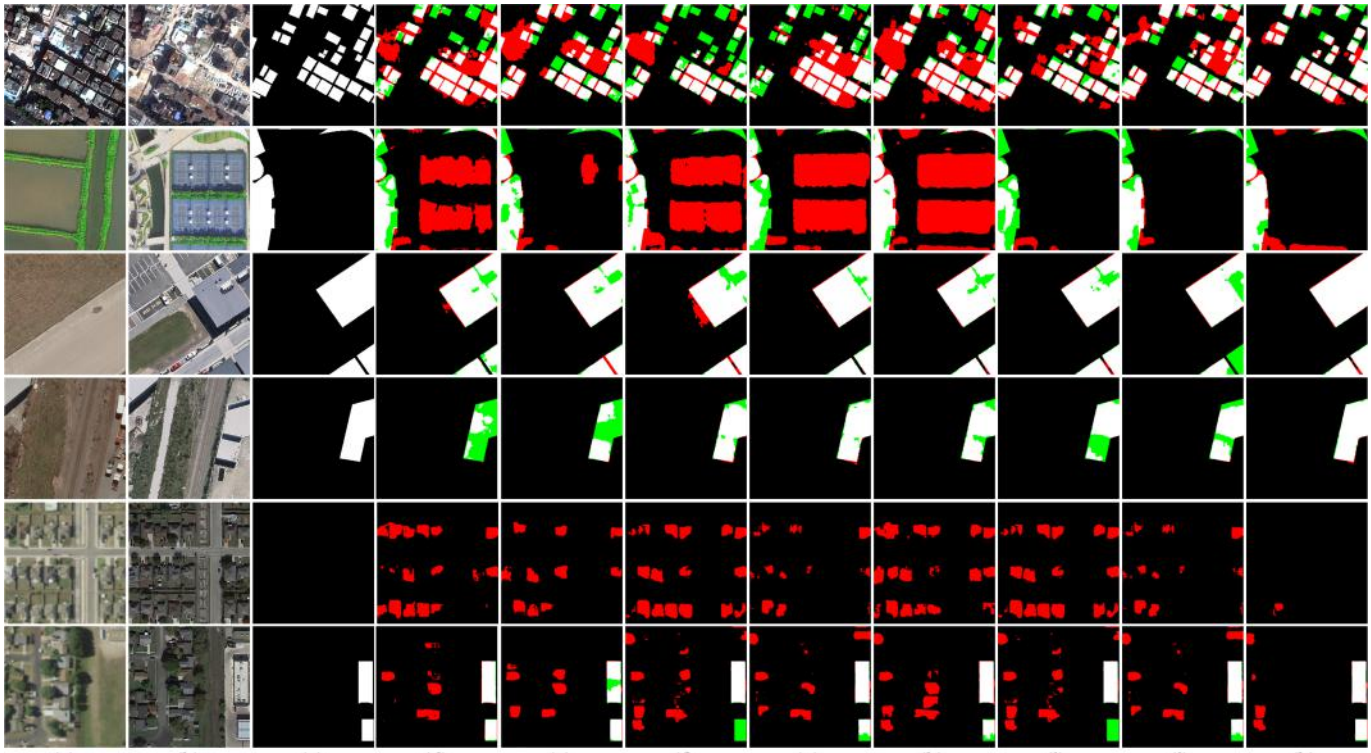


Fig. 9. Visual comparisons of the Ablation experimental model on three datasets. Lines 1 and 2 originate from HRCUS-CD, lines 3 and 4 from WHU-CD, and lines 5 and 6 from LEVIR-CD+. (a) t1 images. (b) t2 images. (c) Ground truth. (d) #1. (e) #2. (f) #3. (g) #4. (h) #5. (i) #6. (j) #7. (k) Ours. The rendered colors represent true positives (white), false positives (red), true negatives (black), and false negatives (green).

comparison with LRRC. Similarly, ablation BaseLines (#4 and #7 on Table III) can be obtained for comparison with LRRC.

To further validate the effectiveness and efficiency of the proposed LDSA module, we conducted a comparative study against two other state-of-the-art feature fusion modules: the adaptive context fusion module (ACFM) from [32] showed in Fig. 5 (b) and the edge-aware refinement module (EARM) from [33] showed in Fig. 5 (c). A core design principle of our decoder is to progressively fuse the deepest features with each shallow feature map individually. To strictly evaluate the fusion module's contribution, we maintained this overarching architecture as a constant for all comparison experiments. We then replaced our LDSA module with the adapted ACFM or EARM modules. As these modules originally output multi-channel features, we appended a 1x1 convolutional layer to their outputs to generate the single-channel attention map required by our decoder, ensuring a consistent comparison framework. The experiments were conducted on the HRCUS-CD dataset, and we evaluated the models based on performance metrics, parameter count, and inference latency.

2) Superiority of CHLR: As shown in Table 4, the addition of LRRC in model #3 and HRRC in model #4 resulted in improvements of 4.68% and 1.35% respectively over the baseline #1 model in the HRCUS-CD dataset. However, when both work together in model #5, there is a 1.66% decrease compared to model #3 with LRRC alone. This decline is due to deep information not being well preserved. After adding LDSA, CHLF-Net shows a 4.78% improvement in F1 over model #5 in the HRCUS-CD dataset. Similar performance improvements are observed in the other two datasets.In Figure 10,

TABLE V
COMPARISON OF DIFFERENT FEATURE FUSION MODULES ON THE HRCUS-CD COLOR CONVENTION: BEST, 2ND-BEST.

K	Latency	Parem.(M)	F1	IoU	Kappa
With ACFM	16.80	1.21	0.7447	0.5932	0.7399
With EARM	21.91	1.33	0.7378	0.5846	0.7331
Our	4.49	1.27	0.7567	0.6086	0.7522

we present several examples of feature visualizations. HRRC effectively captures the core semantics of the target changes of interest, but this may lead to an overemphasis on local targets, resulting in a loss of detailed information. On the other hand, LRRC better captures long-range dependencies, retaining more complete detail information of the target changes of interest, but it may mistakenly identify irrelevant background as change targets. The last column clearly shows that the CHLR effectively integrates core semantic information with long-range dependency for the target changes of interest. By learning deep semantic information from bitemporal images through complementary high-low rank learning, it addresses the current shortcomings in this area.

3) Superiority of LDSA: In the proposed method, LDSA embeds important deep semantic information from the network into the difference features by lightweightly integrating bitemporal features and multi-scale information, which is crucial for the lightweight U-shape network. From the ablation results comparing model #2 and #1 in Table III, it is evident that when using LDSA, the HRCUS-CD, WHU-CD, and LEVIR-

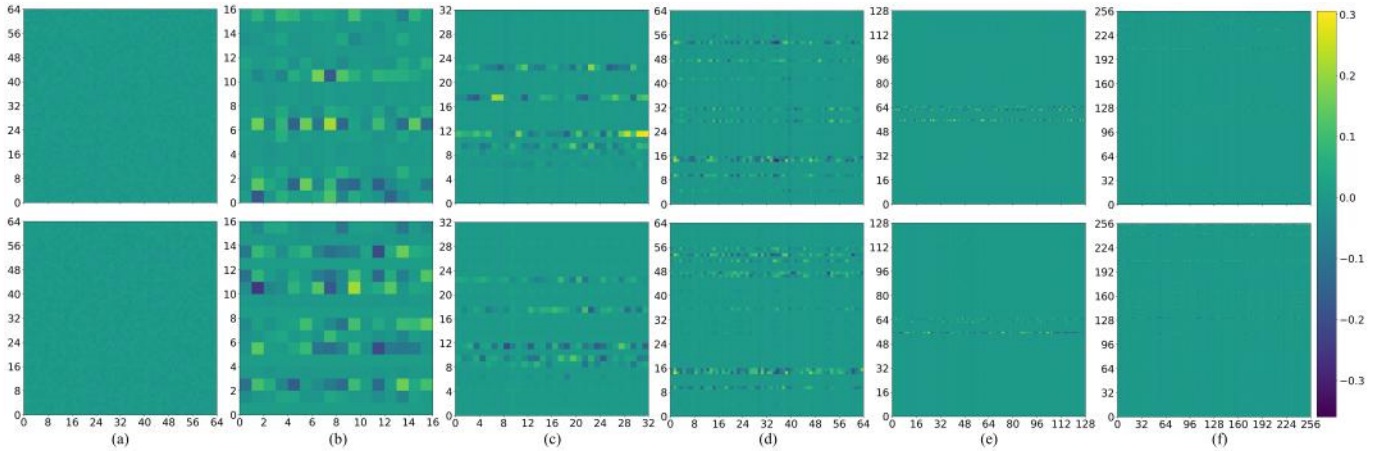


Fig. 10. Visualization of the learned low-rank matrices within our LRRC module for varying ranks (K) on the HRCUS-CD dataset. The two rows correspond to the independently learned matrices for each of the bitemporal images. (a) A randomly initialized matrix (K=64) as a baseline. (b)-(f) Learned matrices for K = 16, 32, 64, 128, and 256. The heatmaps reveal that the model learns structured and increasingly sparse representations.

CD+ improved by 4.90%, 2.14%, and 1.97%, respectively. Comparing model #8 and #5 shows that LDSA also has a significant enhancement effect when using CHLR. From the second row of Fig. 9, with models #2, #6, and #7, although complete targets are not yet segmented, the number of false detections has been greatly reduced. This showcases the superiority of the proposed LDSA.

As presented in Table V, our method equipped with the LDSA module achieves the best performance across all metrics. Specifically, it outperforms the variants using ACFM and EARM in F1-score by 1.20% and 1.89%, respectively. While the model with ACFM has the lowest parameter count, our method is only marginally larger (1.27M vs. 1.21M) yet delivers superior accuracy. Most notably, our approach demonstrates a significant advantage in inference speed, with a latency of only 4.49 ms, which is substantially lower than that of ACFM (16.80 ms) and EARM (21.91 ms). This suggests that while more complex attention or refinement mechanisms can effectively fuse features, they may introduce unnecessary computational overhead and potentially redundant information during the progressive decoding process, especially when the deep features from the encoder are already sufficiently discriminative. The lightweight design of LDSA, in contrast, provides an efficient and effective pathway to integrate deep semantics with shallow differential features, striking an excellent balance between performance and speed.

4) Effect of K: As shown in Table 4, we analyzed the impact of the intrinsic dimension factor K on model performance. As K increases, we observe a performance improvement, with the accuracy peaking at $K = 64$. After this point, a slight decline in performance is observed. We speculate that $K = 64$ is sufficient to capture the critical information in bitemporal images. Additionally, when K is set to 16, 32, and 64, the model's inference time and FPS differ by no more than 0.1 ms and 2, respectively. However, when K increases to 128 and 256, inference speed shows a noticeable rise, increasing by 0.4 ms and 0.9 ms compared to $K = 64$, respectively. Excessive intrinsic dimension factors may introduce more irrelevant semantic information, which not only reduces inference speed but also does not significantly improve performance. Therefore, we chose $K = 64$ as it provides the optimal overall performance.

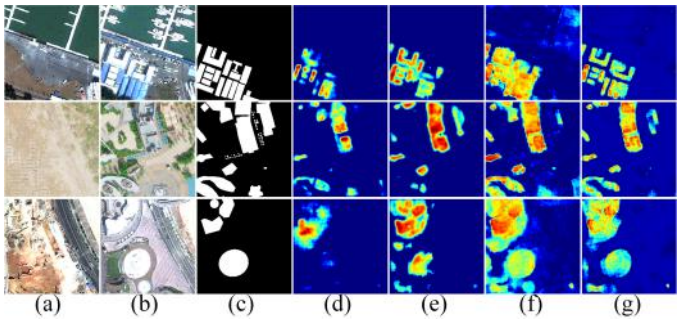


Fig. 11. Visual comparisons of the ablation experimental model on HRCUS-CD. (a) t1 images. (b) t2 images. (c) Ground truth. (d) #1. (e) #6. (f) #7. (g) Ours.

To further understand the working mechanism of the low-rank representation and provide an intuitive explanation for the selection of K, we visualized the learned low-rank matrices from the LRRC module, as shown in Fig. 10. For comparison, a randomly initialized matrix (K=64) exhibits a uniform, noisy distribution, indicating no learned structure. For the trained models, we observe a clear evolution in the learned patterns. At $K = 16$ and 32, the matrices begin to show concentrated features, which become a pronounced and sparse row-wise pattern at $K = 64$ and 128. This suggests that each row has specialized into a distinct semantic basis, efficiently capturing high-level context. Importantly, we note that the two matrices learned for the bitemporal images are themselves distinct, despite sharing similar sparsity characteristics. This demonstrates that the model captures specialized semantic features unique to each temporal image rather than applying a common representation, thereby enabling a more nuanced comparison for change detection. When K is increased to 256, the matrices become denser, aligning with our quantitative results that this may introduce redundancy without improving performance.

V. Discussion

We propose an RSCD method that combines high detection speed with accuracy. The severe spectral changes in bitemporal images and the complexity of the objects of interest present challenges for fast and accurate CD. Our proposed CHLR effectively models contextual information using a relatively lightweight backbone network. LDSA enhances multi-level features with deep features rich in contextual information. Our proposed CHLF-Net generates more accurate predictions with lower computational costs (see Table I and Table II), better detects small change targets, correctly distinguishes irrelevant backgrounds from targets of interest, and segments larger change areas more completely (see Fig. 6, Fig. 7, and Fig.8).

However, there are still some challenging issues that need improvement. In Table III and Fig. 9, we found that when HRRA and LRRC work independently, they do not show significant performance improvement compared to the base model. They must complement each other to accurately capture contextual information. Additionally, the proposed model's segmentation of the edges of change targets is not smooth enough. Nevertheless, extensive experiments demonstrate that the proposed method can handle challenging scenarios well.

VI. Conclusion

This paper present a lightweight, efficient RSCD network, namely CHLF-Net, utilizing centralized high-low rank representation and deep-shallow fusion attention. Our CHLR learns high-level semantics collaboratively through HRRA and LRRC, revealing interesting changes within the complexity of bitemporal HRRS images. HRRA mines high-rank difference features using an MLP structure, while LRRC learns high-level semantics from bitemporal images through low-rank matrices, relating different concepts at a lower computational cost. Then, our proposed LDSA lightweightly integrates deep features rich in contextual information with multi-scale features. Experiments on three RSCD datasets highlight the superiority of the proposed method, which outperforms recent lightweight CD methods, such as TinyCD and A2Net, achieving an effective balance between accuracy and speed.

References

[1] R. E. Kennedy, P. A. Townsend, J. E. Gross, W. B. Cohen, P. Bolstad, Y. Wang, and P. Adams, "Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects," *Remote Sens. Environ.*, vol. 113, no. 7, pp. 1382–1396, 2009.

[2] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, 2015.

[3] Y. Qing, D. Ming, Q. Wen, Q. Weng, L. Xu, Y. Chen, Y. Zhang, and B. Zeng, "Operational earthquake-induced building damage assessment using cnn-based direct remote sensing change detection on superpixel level," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 112, 2022.

[4] 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.

[5] Z. Ying, Z. Tan, W. Li, J. Zhou, Z. Liang, and Y. Zhai, "Uav-bcd: A uav building change detection dataset," in *Dig. Int. Geosci. Remote Sens. Symp. (IGARSS)*. IEEE, 2023, pp. 1012–1015.

[6] D. Wen, X. Huang, F. Bovolo, J. Li, X. Ke, A. Zhang, and J. A. Benediktsson, "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 68–101, 2021.

[7] Z. Lv, T. Liu, J. A. Benediktsson, and N. Falco, "Land cover change detection techniques: Very-high-resolution optical images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 44–63, 2021.

[8] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, 2005.

[9] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[10] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2006.

[11] G. Byrne, P. Crapper, and K. Mayo, "Monitoring land-cover change by principal component analysis of multitemporal landsat data," *Remote Sens. Environ.*, vol. 10, no. 3, pp. 175–184, 1980.

[12] 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, p. 1343, 2019.

[13] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 14, pp. 1194–1206, 2021.

[14] Z. Ying, Z. Tan, Y. Zhai, X. Jia, W. Li, J. Zeng, A. Genovese, V. Piuri, and F. Scotti, "Dgma2-net: A difference-guided multiscale aggregation attention network for remote sensing change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–16, 2024.

[15] 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 Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[16] Y. Wang, L. Gao, D. Hong, J. Sha, L. Liu, B. Zhang, X. Rong, and Y. Zhang, "Mask deeplab: End-to-end image segmentation for change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 104, p. 102582, 2021.

[17] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention siamese network for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–16, 2021.

[18] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–13, 2021.

[19] B. Yang, L. Qin, J. Liu, and X. Liu, "Utrnet: An unsupervised time-distance-guided convolutional recurrent network for change detection in irregularly collected images," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–16, 2022.

[20] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional lstm networks for urban change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 9, pp. 7651–7668, 2021.

[21] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, "A transformer-based siamese network and an open optical dataset for semantic change detection of remote sensing images," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 1506–1525, 2022.

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

[23] W. Li, L. Xue, X. Wang, and G. Li, "Convtransnet: A cnn–transformer network for change detection with multiscale global–local representations," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–15, 2023.

[24] W. Wang, X. Tan, P. Zhang, and X. Wang, "A cbam based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, 2022.

[25] J. Yuan, L. Wang, and S. Cheng, "Stransunet: A siamese transunet-based remote sensing image change detection network," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9241–9253, 2022.

[26] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Wnet: W-shaped hierarchical network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sensing*, 2023.

[27] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, and B. Chen, "Axial cross attention meets cnn: Bibranch fusion network for change detection," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 21–32, 2022.

[28] T. Lei, X. Geng, H. Ning, Z. Lv, M. Gong, Y. Jin, and A. K. Nandi, "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–14, 2023.
[29] A. Codegoni, G. Lombardi, and A. Ferrari, "Tinycd: A (not so) deep learning model for change detection," *Neural Comput. Appl.*, vol. 35, no. 11, pp. 8471–8486, 2023.
[30] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[31] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–11, 2022.
[32] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–14, 2024.
[33] M. Li, D. Ming, L. Xu, D. Dong, and Y. Zhang, "Sfearnet: A network combining semantic flow and edge-aware refinement for highly efficient remote sensing image change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 63, pp. 1–18, 2025.
[34] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022.
[35] Z. Li, C. Tang, X. Liu, W. Zhang, J. Dou, L. Wang, and A. Y. Zomaya, "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–12, 2023.
[36] J. Zhang, Z. Shao, Q. Ding, X. Huang, Y. Wang, X. Zhou, and D. Li, "Aernet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–16, 2023.
[37] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang, "S2looking: A satellite side-looking dataset for building change detection," *Remote Sens.*, vol. 13, no. 24, 2021.
[38] 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, 2018.
[39] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, 2016.
[40] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, 2017.
[41] 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, 2020.
[42] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. Int. Conf. Image Process. (ICIP)*. IEEE, 2018, pp. 4063–4067.
[43] 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, 2020.
[44] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sensing*, vol. 58, no. 10, pp. 7232–7246, 2020.
[45] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Dig. Int. Geosci. Remote Sens. Symp. (IGARSS)*. IEEE, 2022, pp. 207–210.
[46] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–20, 2023.
[47] 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, p. 484, 2020.
[48] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "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, 2020.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[50] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.
[51] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *Int. J. Comput. Vis.*, vol. 131, no. 5, pp. 1141–1162, 2023.
[52] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 9, pp. 7296–7307, 2020.
[53] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Mach. Learn. Res.* PMLR, 2019, pp. 6105–6114.
[54] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. - Int. Conf. 3D Vis. (3DV)*. Ieee, 2016, pp. 565–571.

Yikui Zhai (Senior Member, IEEE) received his Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013. Since October 2007, he has been working with Wuyi University, Jiangmen, China, where he is a Full Professor now. He is also Associate Dean of the School of Electronics and Information Engineering in Wuyi University, since 2021. He has been a Visiting Scholar with Department of Computer Science, the Universit'a degli Studi di Milano, Italy, during June 2016 to June 2017, August 2023 and January 2024.

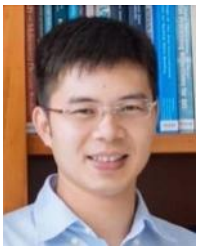
His research interests include: Image Processing, Deep Learning, Optical Character Recognition, Object Detection, UAV Change Detection, SelfSupervise Learning.

Zhanyu Shen received the B.S. degree from Wuyi University, China, in 2025. He is currently pursuing the master s degree with the College of Computer Science and Software Engineering, Shenzhen University.

His research interests include pattern recognition, change detection, large language model.

Junying Zeng (Member, IEEE) received the Ph.D. degree in physical electronics from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008.

He is currently a Professor with Wuyi University, Guangdong, China. His research interests include intelligent signal processing and pattern recognition.



Hongsheng Zhang (Senior Member, IEEE) received the B.Eng. degree in computer science and technology and the M.Eng. degree in computer applications technology from South China Normal University, Guangzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree in earth system and geoinformation science from The Chinese University of Hong Kong, Hong Kong, China, in 2013.

He is currently an Assistant Professor with the Department of Geography, The University of Hong Kong, Hong Kong, China. His research interests include remote sensing applications in tropical and subtropical areas, with a focus on the urban environment and coastal sustainability monitoring, using multisource remote sensing data fusion and image pattern recognition techniques.

Hufei Zhu (Member, IEEE) received the B.E. degree in electical engineering from the University of Science and Technology of China, Hefei, China, in 1998, the M.E. degree in computing from the National University of Singapore, Singapore, in 2004, and the Ph.D. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. He is currently a Research Fellow with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China, and a Distinguished Professor with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, Guangdong, China.

His research activity has led to more than 80 patents granted in China, USA, Europe, Russia, Japan, South Korea, India, and so on, and has also led to numerous publications in leading international journals and conference proceedings. His current research interests include neural networks, signal processing, and wireless communications.

Qingling Chang (Member, IEEE) received the Ph.D. degree from the Institute of Computing Tech nology, Chinese Academy of Sciences, Beijing, China, in 2015.

She is a Master Supervisor and an Associate Professor with Wuyi University, Jiangmen, China, and the Subdecanal of the China-German Artificial Intelligence Institute, Zhuhai, China. She has published more than ten articles included in SCI/EI. Her research interests include artificial intelligence, computer vision, and knowledge graph.

Yan Cui (Member, IEEE) is a Professor with the Faculty of Computer Science, Wuyi University, Jiangmen, China, where he is the Dean of the School of Intelligent Manufacturing, and also the Dean of the China-Germany Artificial Intelligence Institute.

His research interests include computer vision and computer graphic.

Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in industrial and information engineering from the Universita degli Studi della Campania Luigi Vanvitelli, Caserta, Italy, in 2019.

From 2019 to 2022, he was a Post-Doctoral Researcher with the Universita degli Studi di Padova, Padua, Italy. He is currently an Assistant Professor with the Department of Computer Science, Universita degli Studi di Milano, Milan, Italy. His research interests include theoretical, methodological, and applied aspects of computational intelligence for signal and image processing.

Dr. Coscia has been the Co-Chair of the Intelligent Measurement Systems Technical Committee (TC-22) of the IEEE Instrumentation and Measurement Society since 2023.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer engineering and automation from the Politecnico di Milano, Milano, Italy.

He is currently a Full Professor with the Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters. His main research interests include pattern recognition, machine learning, machine perception, robotics, computer vision, signal and image processing, three-dimensional reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Dr. Genovese is an Associate Editor of the Journal of Ambient Intelligence and Humanized Computing (Springer).