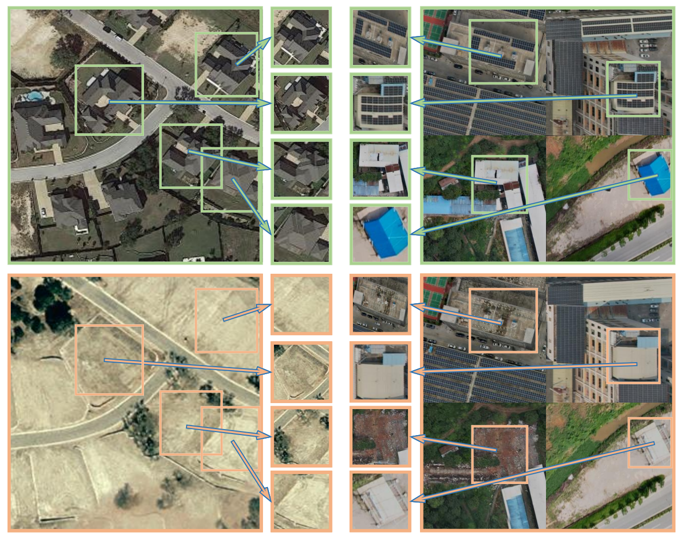
This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3502768

Efficient Adjacent Feature Harmonizer Network with UAV-CD+ Dataset for Remote Sensing Change Detection

Yikui Zhai[1*], *Senior Member, IEEE*, Jiahao Pan[1*], *Student Member, IEEE*, Hongsheng Zhang[2], *Senior Member, IEEE*, Tingfeng Xian[1*], *Student Member, IEEE*, Ying Xu[1*], Pasquale Coscia[3], *Senior Member, IEEE*, Angelo Genovese[3], *Senior Member, IEEE*, Vincenzo Piuri[3], *Fellow, IEEE*, Fabio Scotti[3], *Senior Member, IEEE*, C. L. Philip Chen[4], *Fellow, IEEE*.

***Abstract*—Remote sensing change detection (RSCD) aims to identify changes within bi-temporal registered images. However, existing deep learning-based RSCD networks often suffer from large numbers of parameters, high computational complexity, and low inference speed, making it challenging to achieve efficient inference in real-world deployments. Additionally, current models lack robust feature fitting capabilities, necessitating the development of an efficient and powerful RSCD model to address this issue. Therefore, we propose a novel RSCD network named Efficient Adjacent Feature Harmonizer Network (EAFH-Net) with fast computational speed and lightweight design. It is based on MobileNetV2, considering that change maps of different sizes contain temporal information of bitemporal features and spatial information at various scales, we introduce Multiscale Feature Neighbor Fusion Module (MFNFM) to address the lack of interaction between sophisticated-level and elementary-level features, and Spatial and Channel Harmonizer Module (SCFHM) to harmonize the spatiotemporal information of the change maps. Moreover, data-driven deep learning (DL) algorithms face another challenge due to insufficient granularity and the need for more practical datasets. Therefore, we present UAV-CD+, a dataset comprising 2002 pairs of bi-temporal UAV low-altitude images, each sized at 1024×1024. We performed experiments on three publicly accessible datasets in conjunction with UAV-CD+, comparing the results with other state-of-the-art (SOTA) methods. EAFH-Net attains the utmost precision, obtaining 91.74% on LEVIR-CD, 84.28% on SYSU-CD, 95.07% on WHU-CD, 79.12% on CLCD, and 70.12% on UAV-CD+. We have our model code available at the following link: https://github.com/yikuizhai/UCSFH-Net.**

***Index Terms*—Remote sensing change detection, Ultra-lightweight, UAV Datasets, multiscale feature neighbor fusion module, spatial and channel feature harmonizer module.**



(a) (b)

Fig. 1. Samples of two images taken at different points in time. (a) Satellite imagery; (b) UAV imagery.

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011576), Guangdong Science and Technology Planning Project (No. 2021A0505030080, No. 2021A0505060011), Guangdong Higher Education Innovation and Strengthening School Project (No. 2020ZDZX3031, No. 2022ZDZX1032, No. 2023ZDZX1029), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology Research Project (No. 2220002000246, No. 2023760300070008390). *(Corresponding author: Yikui Zhai and Hongsheng Zhang.)*

Yikui Zhai, Jiahao Pan, Tingfeng Xian, Ying Xu are with the College of Electronics and Information Engineering, Wuyi University, Jiangmen, 529020, China (e-mail: yikuizhai@163.com; jiahaopan88@163.com; tingfengxian21@163.com; xuying117@163.com).

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

Pasquale Coscia, Angelo Genovese, Vincenzo Piuri and Fabio Scotti are with the Department of Computer Science and Dipartimento di Informatica, Universita Degli Studi di Milano, 20133, Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it)

C. L. P. Chen is the Faculty of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510006, China (e-mail: philip.chen@ieee.org).

*Contributed to this work equally.

I. Introduction

REMOTE sensing change detection (RSCD) is a methodology that identifies and segments changing parts within the same geographical area across different time intervals by collecting registered multi-temporal remote sensing (RS) images. RSCD is utilized across diverse domains such as urban planning [1], land cover change detection (CD) [2], catastrophe evaluation [3], forest cover change [4], resource management [5], and environmental monitoring [6]. In the past, due to the limitations of RS satellites as data collection platforms, including insufficient spatial resolution that made it difficult to capture very small changes in ground objects, long observation intervals that hindered the collection of high-frequency multi-temporal images, and high costs and technical barriers, multi-temporal RSCD faced significant challenges. On the other hand, the high spatial resolution, flexible data collection, and low cost of unmanned aerial vehicle (UAV) make them a promising future mainstream platform for RSCD data collection.

Previously, early CD tasks depended on traditional handcrafted methods for feature extraction, Examples of these methods include Image Differencing (ID) [7], Principal Component Analysis (PCA) [8], and Change Vector Analysis (CVA) [9]. However, achieving high accuracy in CD tasks through traditional handcrafted feature extraction methods proved to be challenging. In recent years, with the rapid advancement of Convolutional Neural Networks (CNNs) and deep learning (DL), the robust feature extraction capability of CNNs has garnered significant attention and adoption among researchers. Downstream tasks such as image segmentation [10], [11] and object detection [12], [13] have demonstrated improved performance through CNN-based feature extraction.

Currently, many CD methods rely on feature extraction through CNNs. However, according to experiments and statistics, existing CD models with good accuracy suffer from high computational costs and large model sizes, making it challenging to deploy them on low-computational-edge devices. To facilitate the deployment of CD models on UAVs and save costs associated with transmitting bi-temporal RS images captured by UAVs to end devices for inference, there is a need to develop high-precision, fast-inference, and lightweight models for this task. While previous CD models have shown excellent performance only in bi-temporal images captured by RS satellites, to achieve high-precision CD in both satellite and UAV images, we propose the Ultralightweight CD network with comparable parameter count and lower computational costs called Efficient Adjacent Feature Harmonizer Network (EAFH-Net). Conversely, taking into account the limitations inherent in historical satellite imagery and high-altitude aerial imagery, particularly their low spatial resolution,which hinders the capture of subtle changes on the Earth's surface, we introduce a novel low-altitude UAV aerial CD dataset named UAV-CD+. We will conduct experiments using three publicly available satellite and aerial CD datasets, as well as our proposed UAV-CD+ dataset. The results will demonstrate that our proposed EAFH-Net achieves state-of-the-art (SOTA) performance, offering superior results with reduced parameter count and computational cost.To achieve the lightweight design of the CD network, we utilize MobileNetV2 [14] as the backbone network instead of the ResNet series [15]. The satellite-captured images and UAV-captured images are shown in Fig. 1.

In CD tasks, MobileNetV2 can serve as a lighter contextual encoder compared to ResNet, with performance comparable to or even superior to ResNet. Inspired by multi-scale feature fusion methods like Feature Pyramid Networks (FPN) [16] and Pyramid Attention Networks (PAN) [17], we devised a Multiscale Feature Neighbor Fusion Module (MFNFM) to incorporate features acquired by the backbone network. This module adeptly merges sophisticated and elementary-levels features extracted by MobileNetV2, strengthening the extracted bi-temporal information and refining granularity to achieve better performance.

In addition to the mentioned encoders, we designed a computationally efficient and lightweight decoder, which includes a Spatial and Channel Feature Harmonizer Module (SCFHM). The SCFHM acts as a harmonizer for encoding the difference maps. It harmoniously integrates the bi-temporal features of the difference maps across spatial and channel dimensions, making them more prominent in the feature space. By harmonizing both sophisticated and elementary-levels features and synergistically blending them across spatial and channel dimensions, it facilitates a better integration of the bi-temporal features, thereby presenting the features more comprehensively and enhancing the precision of the CD model's detections.

By combining MFNFM and SCFHM, our proposed network EAFH-Net achieves F1-scores of 91.74%, 84.24%, 95.07%, 79.21%, and 70.12% on the LEVIR [18], SYSU [19], WHU-CD [20], CLCD [21] and our proposed UAV-CD dataset UAV-CD+, respectively. Furthermore, EAFH-Net has a small parameter count (4.81M) and extremely low FLOPs (2.93G), making it the CD model with excellent accuracy and minimal FLOPs known to date. Below, Our contributions can be encapsulated as:

1) We introduce an innovative Siamese-Network framework named EAFH-Net, which consists of MFNFM and SCFHM modules. The MFNFM module facilitates the cross-fusion of sophisticated and elementary-levels in proximity to the backbone network, while the SCFHM module harmonizes spatiotemporal feature information of the change map output by the encoder, further amalgamating sophisticated and elementary-levels features to ensure the comprehensiveness of difference map features.

2) Harnessing a lightweight backbone network streamlines the extraction of bi-temporal features. Furthermore, the integration of MFNFM and SCFHM, which embed efficient local convolution processing, enables effective interaction across multi-scale features and enriches spatiotemporal information within the change map. This strategy culminates in achieving state-of-the-art performance with a modest parameter count of 4.81M and computational efficiency of 2.93G Flops.

3) We propose UAV-CD+, a UAV-based CD dataset comprising 2002 pairs of 0.06 m-resolution RSCD images, each boasting a resolution of 1024x1024 pixels. This dataset covers urban building changes, land changes, water body changes, pre-construction and post-construction changes in construction sites, enabling a broader range of CD tasks.

4) We undertake experimentation across four benchmark datasets. Furthermore, we include our devised dataset to ascertain the efficacy and supremacy of the proposed technique. The outcomes demonstrate that EAFH-Net outperforms twelve SOTA models in both accuracy and efficiency.

The subsequent sections of this article are organized as outlined below: Section II offers a concise overview of relevant research; Section III delves into the intricate nuances of the novel UAV-CD+ dataset; Section IV unveils the architectural framework of the EAFH-Net; Experimental intricacies are

TABLE I
DIFFERENT CD DATASET INFORMATION

Dataset	Number of Image Pairs	Image Size	Resolution(M)	Modality	Platform/Source
OSCD [22]	24	600×600	10	Multispectral	Sentinel-2 satellites
CDD [23]	16000	256×256	0.03-1	RGB	Google Earth(satellites)
S2Looking [24]	5000	1024×1024	0.5-0.8	RGB	GaoFen,SuperView, BeiJing-2 (satellites)
LEVIR-CD [18]	637	1024×1024	0.5	RGB	Google Earth(satellites)
SYSU-CD [19]	20000	256×256	0.5	RGB	aerial image
WHU-CD [20]	1	32207×15354	0.2	RGB	aerial image
CLCD [21]	600	512×512	0.5-2	RGB	aerial image
DSIFN-CD [25]	3940	512×512	0.03-1	RGB	Google Earth(satellites)
UAV-CD+(ours)	**2002**	**1024×1024**	**0.06**	**RGB**	UAVs

elaborated upon in Section V, and lastly, conclusions are encapsulated in Section VI.

II. RELATED WORK

A. Traditional CD Methods

Early CD methods traditionally depended on hand-crafted techniques for extracting features from bi-temporal images. For instance, methods like Image Differencing [7] involved subtracting images from two different time points to generate a difference map that depicts changes occurring on the Earth's surface. Subsequently, thresholding was applied to segment the difference map to extract and display the changed areas. Additionally, PCA [8] utilizes linear transformations to condense the dimensionality of the original data, retaining the maximum data variance in the lower-dimensional space. In the context of CD tasks, PCA transformed bi-temporal images into lower-dimensional information, with selected thresholds determining principal components and changed components. CVA [9] computed pixel-level differences between bi-temporal images, yielding change vectors comprising the magnitude and direction of change for each pixel between the two time points. Finally, statistical analysis of these change vectors, along with thresholding, determined significant changes. These methods primarily relied on hand-crafted feature extraction, resulting in lower accuracy and insufficient robustness. Furthermore, in the current landscape, characterized by satellite imagery with exceptional resolution or UAV imagery, these methods encounter challenges in excelling at CD tasks.

B. DL-Driven CD Approaches

With the meteoric progress of DL, many CD methods based on CNNs and Transformers [26]–[31], have been proposed by researchers. These methods have shown superior accuracy and robustness compared to traditional CD approaches. CNNs are utilized for local feature extraction from bi-temporal images, where sophisticated level features extracted can locate changed objects, while elementary level features capture intricate changes within the altered regions. Moreover, modules like FPN and PAN have emerged in the DL field to integrate sophisticated and elementary levels features. Researchers have extensively crafted networks to amalgamate the sophisticated and elementary levels features gleaned by CNNs, aiming to attain heightened precision in CD. For instance, early Fully Convolutional Network (FCN) [32]and its variants are designed for fine-grained CD tasks. Daudt et al. [33] proposed three FCN structures, including the FC Encoding and Decoding Network with skip connections (FC-EF) and two Siamese FCNs (FC-Siam-conc and FC-Siam-diff). Barkur et al. [34] introduced RSCDNet, a pixel-level CD model tailored for bi-temporal high-resolution (HR) RS images. The model integrates the MSA and GL-ASPP blocks. Meanwhile, a plethora of weakly supervised learning and unsupervised learning methods have surfaced. Zhao et al. [35] introduced a CD network named FMCD, informed by feature interplay and multi-pronged learning. The contextual information features of the FMCD model are acquired through multi-tier feature interaction modules to extract distinctive features. WU et al. [36] presented a FCN CD framework employing generative adversarial networks (GAN). This framework amalgamates unsupervised, weakly supervised, region supervised, and fully supervised CD tasks into a unified structure. subsequently, utilizing the transformer decoder to reintegrate the enriched tokens into the pixel space, thereby enhancing the original features. However, past CD methods often utilized large-parameter backbones and complex feature fusion modules [37]–[40], leading to models with excessive parameters and high computational costs. This made it challenging to deploy them for inference on edge devices, such as UAVs, due to their high computational demands. Therefore, we lean towards designing a CD model with extremely fast computation speed, minimal parameters, and high accuracy.

C. Lightweight CD methods

In the era of ubiquitous artificial intelligence, where DL models are routinely deployed on mobile devices, researchers increasingly pursue DL models with lower parameter counts and reduced computational costs. This pursuit extends to tasks like RSCD, which also require more efficient lightweight networks.

Li et al. [41] proposed a novel lightweight network named A2Net, which integrates features from adjacent stages through a neighbor aggregation module. They also introduced a progressive change identifying module (PCIM) to extract tem-

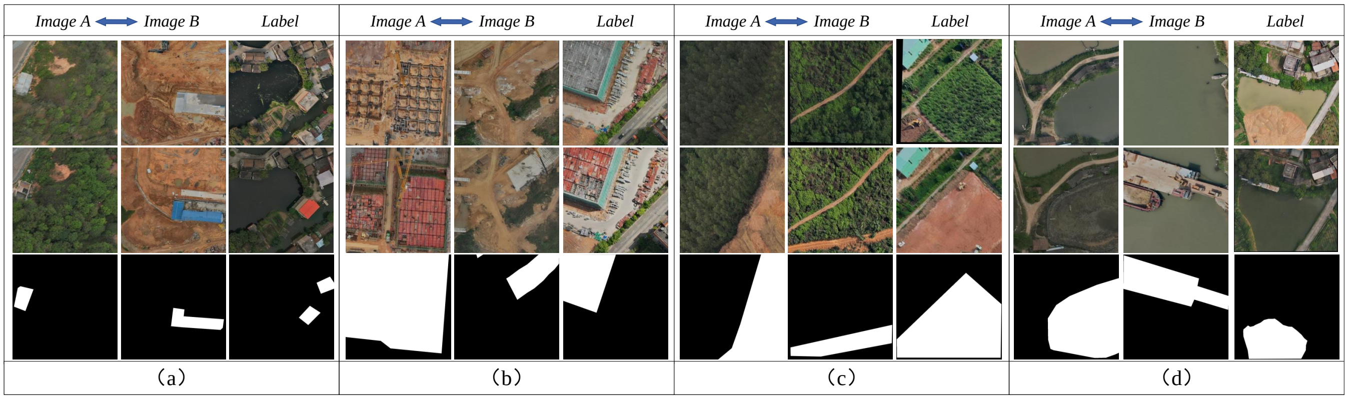


Fig. 2. Examples of captured regions (1024×1024) from our UAV-CD+ dataset. Each column represents a category of change: (a) Changes in urban architectures; (b) Changes in large-scale construction sites; (c) Land changee; (d) Changes in water areas.

poral difference information from bitemporal features, along with a supervised attention module (SAM) for effective reweighting between high-level and low-level features. Lei et al. [42] introduce an ultralightweight spatial-spectral feature cooperation network named USSFC-Net. The network's strength lies in its multiscale decoupled convolution (MSDConv) design, which flexibly captures multiscale features of changing objects. This design significantly reduces parameter count and computational redundancy. Additionally, the network introduces spatial-spectral feature cooperation (SSFC) to enrich its capabilities. Li et al. [43] proposed a lightweight network that uses a focus module for downsampling in the encoding layer. They designed an efficient feature extraction module based on depthwise convolutions by stacking convolutional kernels of different sizes, and introduced a low-parameter multi-scale feature fusion module. Jiang et al. [44] proposed a full-scale hybrid network that combines CNN, MLP, and Transformer. The MLP helps recover lost global information, while the Transformer effectively establishes the association between semantic information and pixel space, enabling more efficient structure for CD tasks. Recently, many lightweight CD networks have also been proposed [45], [46].

Some of the current lightweight CD models still adopt feature fusion methods with high computational complexity or overly simplistic architectures, making it difficult to ensure both model efficiency and performance. With the increasing prevalence of lightweight devices, there is a need for a model that maintains a lightweight structure while achieving high accuracy in CD tasks across various scenarios. Such a model should be deployable on lightweight devices, enabling fast, efficient real-time inference. To ensure the network remains lightweight, we adopt MobileNetV2 as backbone, with MFNFM for cross-fusion of adjacent sophisticated and elementary levels features after features extraction. The decoder includes SCFHM which acts as a harmonizer for encoding the difference maps. It harmonizes and integrates bi-temporal features from difference maps across spatial and channel dimensions, enhancing their prominence in the feature space. This harmonization and blending of both sophisticated and elementary levels features facilitates a more comprehensive integration of bi-temporal features, ensuring a more thorough representation.

III. UAV-CD+ DATASET

In the course of CD development, considerable efforts have been dedicated to the creation of CD datasets (CDDs), which play an essential role in advancing the CD field. Table I presents the collected CDDs image pairs, their sizes, resolutions, and modal types, including those captured by satellites [47] and UAVs [48]. For example, the OSCD dataset was established using Sentinel-2 satellite imagery, capturing images at resolutions ranging from 10 meters to 60 meters across 13 bands between ultraviolet and shortwave infrared. This dataset comprises 24 pairs of 600×600 pixel images and has been one of the significant datasets in early multispectral CD research.

Following this, in order to facilitate high resolution change detection (HRCD) tasks, researchers introduced the CDD dataset, which involved cropping images from 7 pairs of real seasonal change RS images. These images were sized at 4725×2700 pixels, downscaled to smaller image pairs formatted in dimensions of 256 by 256 pixels and spatial resolutions spanning from 0.03m to 1m. Additionally, for instance-level CD tasks, the S2Looking dataset was introduced, comprising 5000 pairs of large-scale oblique satellite images captured at different non-nadir angles. This dataset included over 65290 annotated change instances from rural areas.

Nowadays, many datasets proposed by researchers are specifically designed for monitoring and detecting urban building changes. The most commonly used datasets include LEVIR-CD [18], SYSU-CD [19], WHU-CD [20], CLCD [21], and DSIFN-CD [25]. Compared to other datasets, our proposed UAV-CD+ dataset features diverse categories of change objects and higher resolutions, aiming to enable CD models to achieve better performance in practical CD applications.

While the datasets mentioned above tackle the issue of CD in multi-temporal RS images, it's worth noting that most bi-temporal data are captured by RS satellites or high-altitude aerial platforms. The image resolution and spatial resolution of these datasets are often suitable for CD tasks involving significant changes but may not be sufficient for tasks involving subtle changes or small change areas. Additionally, many

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3502768

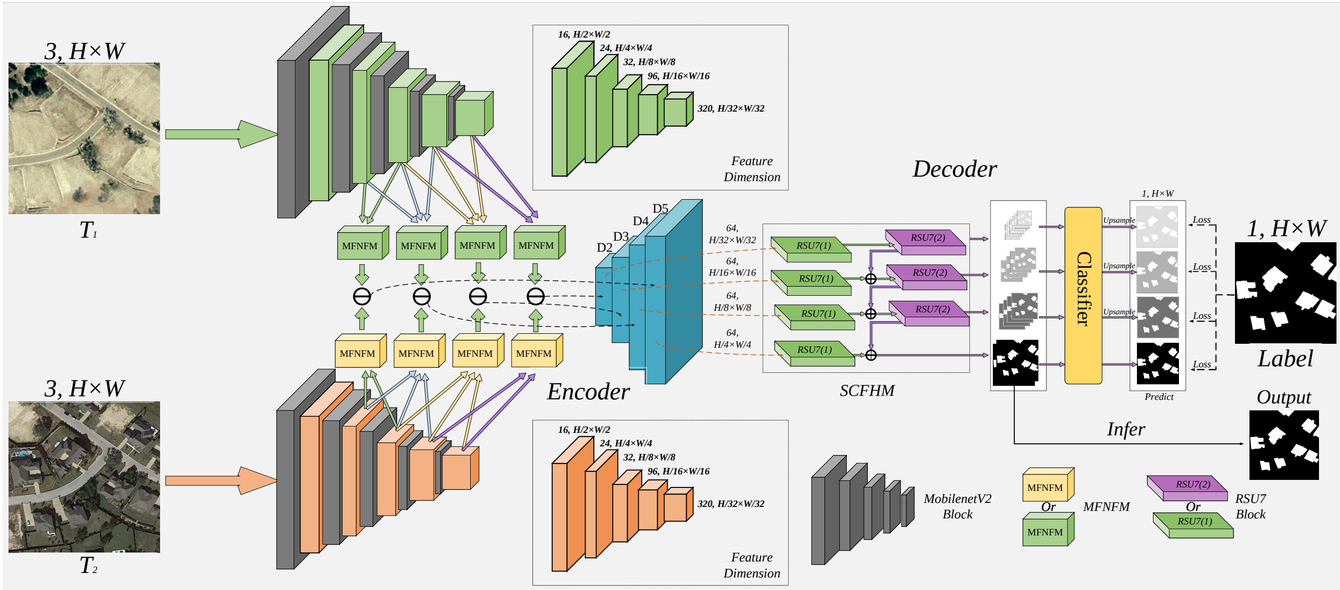


Fig. 3. The proposed EAFH-Net framework utilizes the MobileNetV2 network as backbone in a Siamese-Network configuration with weight sharing, to extract features of bi-temporal image pairs. Features from different scales are fed to the MFNFM module for the spatial information fusion of adjacent features. The output difference map undergoes spatial and channel feature information reconciliation through the SCFHM module, which consists of two stacked Residual U-blocks-7 (RSU7) layers. This process further blends sophisticated and elementary levels features. Finally, scale alignment is performed through upsampling, yielding the change map as the output.

datasets focus exclusively on urban building changes, limiting their applicability to non-building CD tasks and failing to meet the requirements of practical applications.

Our proposed UAV-CD+ dataset utilizes low-altitude drone imagery to complement existing urban building CD datasets. It addresses the need for detecting subtle changes and non-urban building changes, providing a new benchmark for CD and UAV-based CD tasks. The dataset was captured in a prefecture-level city in Guangdong Province and consists of 2002 pairs of images captured by low-altitude drones. Such as other CD datasets, our proposed dataset is suitable for evaluating the performance of CD models in UAV-based change detection tasks. Traditional satellite remote sensing images and high-altitude aerial images suffer from issues such as insufficient resolution, difficulty in data acquisition, complex processing procedures, and susceptibility to meteorological conditions. Furthermore, some large-scale RSCD datasets still have room for improvement. For instance, the spatial resolution of certain datasets remains low, and some datasets are relatively small, which may lead to overfitting in data-driven deep learning techniques. As shown in Table I, the dataset we propose, collected via UAV, has the highest spatial resolution. Moreover, for images of the same size, our dataset surpasses the SYSU-CD, LEVIR-CD, and WHU-CD datasets in terms of the balance between spatial resolution and dataset size. Each image is dimensioned at 1024×1024, boasting a spatial resolution of 0.06m, and there is a two-year interval between captures. Due to the rapid urban development in Guangdong Province, there is a frequent occurrence of illegal construction and unauthorized land occupation. We not only collect data on rapid changes in basic infrastructure, but also gather information on changes in land use, certain large-scale construction sites, and alterations in water bodies. Therefore, our dataset includes a large number of instances of illegal land occupation, captured through low-altitude drone flights to detect instances of small-scale illegal occupation more accurately. We partitioned the initial 2002 pairs of images into training, validation, and test sets, maintaining an 8:1:1 ratio. To address device performance limitations, We cropped the images into numerous pairs, each with a pixel dimension of 256×256. During training, we applied random rotation and cropping for data augmentation. The predominant changes in the datasets are visually depicted in Fig. 2, including: (a) architectural changes, (b) large-scale construction sites, (c) land alterations, and (d) changes in water areas.

IV. Methodology

In this section, we commence by presenting the proposed model architecture. Subsequently, we offer an intricate breakdown of each constituent of our model. Lastly, we introduce the optimization strategy employed for the network model.

A. Overview

Our proposed architecture, depicted in Fig. 3, comprises three components: the backbone network, MFNFM module, and the decoder. We employ a Siamese network with weight sharing to extract features from bi-temporal images. Prior research has shown that utilizing a backbone network with pre-trained parameters aids in model convergence. Therefore, for real-world deployment scenarios, we select MobileNetV2 as the backbone extractor, leveraging pre-training on the ImageNet dataset.

To acquire deep features in CD, the backbone feature extraction network extracts multi-layered feature information. These features are subsequently input into the MFNFM for the aggregation of multi-scale adjacent features. The aggregated features are then passed into the SCFHM module in the decoder to harmonize different channel information.

If we designate image pairs taken at distinct time intervals within the identical location as T_1 and T_2 respectively, and denote the labeled ground truth with change area annotations as L, the comprehensive procedure of EAFH-Net unfolds as outlined below:

1) Initially, the images input T_1 and T_2 undergo feature extraction via the backbone network, which utilizes weight sharing. This results in a collection of multi-scale features, denoted as $\{feat_1^t, feat_2^t, feat_3^t, feat_4^t, feat_5^t\}$, where t indicates the time points T_1 and T_2.

2) Next, the multi-scale features extracted from the shared-weight dual-branch backbone feature extractor are cross-fed into our proposed MFNFM. The features extracted by the dual-branch feature extractor are $\{feat_2^{t1}, feat_3^{t1}, feat_4^{t1}, feat_5^{t1}\}$ and $\{feat_2^{t2}, feat_3^{t2}, feat_4^{t2}, feat_5^{t2}\}$. $MF(\cdot)$ refers to the operation of features entering the MFNFM.

$$P_n^t = MF(feat_n^t) \quad n \in (2,3,4,5) \quad t \in (1,2) \tag{1}$$

This process enables fusion of channels and scales between adjacent features extracted from the backbone feature extractor. Following this, the features from both branches, post MFNFM processing, undergo an absolute difference operation as outlined below:

$$D_i = \left|p_n^{t1} - p_n^{t2}\right| \quad n \in (2,3,4,5) \tag{2}$$

3) This operation produces the CD difference maps as shown in Eq. 2. Then, the differentially encoded features from both branches are input into the decoder for decoding operations. Then, the features maps D_2, D_3, D_4, D_5 from different scales are input into the Spatial and Channel Feature Harmonizer Module (SCFHM), resulting in $\{D_2^{'}, D_3^{'}, D_4^{'}, D_5^{'}\}$.Given the necessity to reconcile lower-level features with spatial and channel information, the integrated lower-level features undergo upsampling and subsequent fusion with adjacent higher-level features. This process aims to achieve better fusion of feature information between different scales, enhancing the change features to acquire more precise change data. Here, $SC(\cdot)$ signifies the feature difference map following the operation process of SCFHM. After the SCFHM self-enhancement of the features of the change map, four high-dimensional change feature maps of different scales are obtained. After passing through the second layer RSU block of SCFHM, these four feature maps undergo dimensionality reduction through a classifier, resulting in four different scale change results with a channel number of 1.

$$D_i^{'} = SC(D_i) \quad i \in (2,3,4,5) \tag{3}$$

4) Finally, the change results from the four scales are upsampled to a pixel size of 256×256, C_5 is the change map output by the inference mode. resulting in change results C_i, where i belongs to $\{2,3,4,5\}$. These change results are then used to compute the final loss (see Section IV.D). The sum of these metrics across all scales promotes more accurate model training.

$$C_i = Upsampling(D_i^{'}) \quad i \in (2,3,4,5) \tag{4}$$

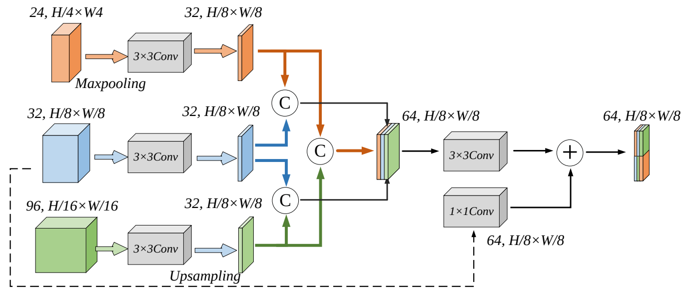


Fig. 4. Structure of the proposed MFNFM.

B. Multiscale Feature Neighbor Fusion Module

Considering the independence of features extracted from the backbone, isolated sophisticated or elementary information cannot simultaneously address both coarse-grained and fine-grained features. This is detrimental to the precise localization and accurate pixel value prediction required for the final change detection task. Performing convolution operations on feature maps of varying sizes in a simplistic and independent manner fails to facilitate the integration of features across different levels. Given that each layer's features contain specific representations, it is essential to consider these unique representations within each layer. Therefore, we concatenate adjacent multiscale features in pairs to enable the integration of their specific representations. This approach enhances the completeness of the adjacent multiscale feature representations obtained from MFNFM. Our subsequent experiments demonstrate the effectiveness of this concatenation operation.

$$F_2^1 = Conv_{3\times3}(Maxpooling(feat_2)) \tag{5}$$

$$F_3^1 = Conv_{3\times3}(feat_3) \tag{6}$$

$$F_4^1 = Upsampling\{Conv_{3\times3}(feat_4)\} \tag{7}$$

Describing MFNFM in detail, this module is designed to facilitate interaction among adjacent features. For instance, features $feat_2$, $feat_3$, and $feat_4$ extracted from the backbone network undergo the following operations to enrich channel information and refine features: $feat_2$ is subjected to max-pooling downsampling and a 3×3 convolution to increase channel dimensions, resulting in F_2^1; $feat_3$ undergoes further feature refinement via a direct 3×3 convolution without altering channel dimensions, yielding F_3^1; finally, $feat_4$, with high channel dimensions, undergoes channel reduction and upsampling to focus more on fine-grained changes, resulting in F_4^1. Subsequently, the three aligned feature maps $\{F_2^1, F_3^1, F_4^1\}$ are concatenated pairwise along the channel dimension, allowing adjacent features to interact and learn common characteristics, optimizing the edges and textures of change areas. This enhances the network's ability to delineate change areas, resulting in feature maps $\{F_2^2, F_3^2, F_4^2\}$, which are then concatenated along the channel dimension to produce fully integrated features. These features are subsequently

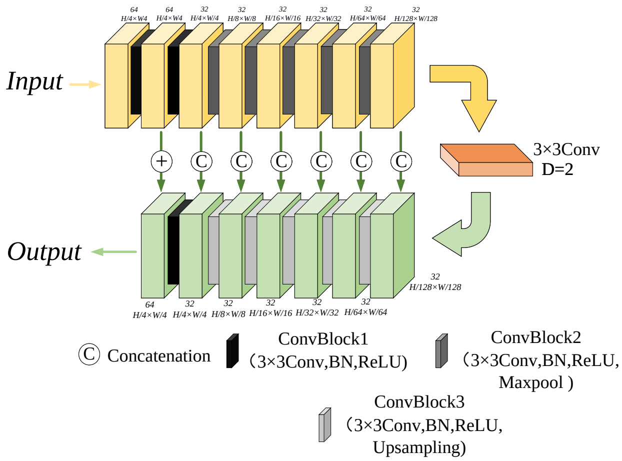


Fig. 5. The overview of Residual U-block.

dimensionally reduced using a 3×3 convolutional block to acquire the final integrated features.

$$F_2^2 = Concat(F_2^1, F_3^1) \tag{8}$$

$$F_3^2 = Concat(F_3^1, F_4^1) \tag{9}$$

$$F_4^2 = Concat(F_2^1, F_4^1) \tag{10}$$

Additionally, similar to residual operations, F_3 undergoes channel dimensionality increase through a 1×1 convolution to embed original information into the integrated features, preventing them from deviating too much from the original features. This ensures the model predicts change areas more accurately and with finer detail. The bi-temporal image undergoes this step to obtain $\{P_2(t), P_3(t), P_4(t), P_5(t)\}, t \in (1, 2)$. The specific structure is illustrated in Fig. 4.

$$P_3(t) = Conv_{1\times1}(feat_3) + Conv_{3\times3}(Concat(F_2^2, F_3^2, F_4^2)) \tag{11}$$

C. Spatial and Channel Feature Harmonizer Module

Considering that the difference maps D_2, D_3, D_4, D_5 obtained through the Encoder only provide coarse-grained change location information and struggle with precise pixel-level predictions, a deeper feature extraction method is needed to enhance the edge and texture information of the difference maps. To address this, we employ two layers of RSU-blocks [49] in SCFHM to deeply reinforce the precise change information within the difference maps themselves. While RSU-blocks delve into deep information mining of the difference maps, residual operations within the RSU-blocks integrate shallow and deep information of the difference maps, preventing excessive smoothing of fine-grained features. This approach enables simultaneous and accurate prediction of changed regions and their precise changed pixels, the following RSU operations refer to calculations performed through the Residual U-block. The structure is illustrated below:

$$D_5^{''} = RSU\left[RSU(D_5)\right] \tag{12}$$

$$D_i^{''} = RSU\left(D_{i+1}^{''} + RSU(D_i)\right) \quad i \in (3,4) \tag{13}$$

$$D_i^{''} = D_{i+1}^{''} + RSU(D_i) \quad i = 2 \tag{14}$$

The RSU conducts multi-layer 3×3 convolutions for feature extraction along with max-pooling downsampling operations to better capture the global features and structural information of the image. Additionally, it utilizes a 3×3 convolution block with a dilation of 2 in the final stage to enhance the receptive field for feature extraction while reducing computational overhead, The $DConv_{3\times3}$ in Eq.16 refers to the above dilated convolution operation. Following that, via continuous upsampling and concatenation with features from the downsampling stage, sophisticated and elementary levels features are combined to maintain the integrity of feature information. This particular structure is illustrated in Fig. 5.

$$D_{2(f)}^{j=7} = \left\{ConvBlock2\left[ConvBlock1(D_2)\right]^{(k=2)}\right\}^{(l=5)} \tag{15}$$

$$D_{2(b)}^{j=7} = DConv_{3\times3}(D_{2(f)}^{j=7}) \tag{16}$$

$$D_{2(b)}^{(j-1)} = ConvBlock3\left[Concat\left(D_{2(b)}^{j}, D_{2(f)}^{j}\right)\right] \tag{17}$$

$$D_2^{'} = ConvBlock1\left(D_2\right) + ConvBlock1\left(D_{2(b)}^2\right) \tag{18}$$

The above Equations Eq.15-Eq.18 represent the computation operations in the RSU, where k in Eq.15 denotes the 2-layer $ConvBlock1$ operations, and l denotes the 5-layer $ConvBlock2$ operations. The specific $ConvBlock1$, $ConvBlock2$ and $ConvBlock3$ operations are shown in Fig. 5. And j represents the level of feature hierarchies in the RSU's forward process, where $3 < j < 8$ and j is an integer; f and b refer to the forward and backward computation operations in the RSU, respectively.

D. Loss function

The loss function utilized comprises a amalgamation of Binary Cross-Entropy Loss (BCELoss) and Dice Loss [57]. Since CD focuses more on predicting foreground objects, and typically foreground objects occupy a small proportion of the image, Dice Loss assists in mitigating the adverse effects of foreground-background class imbalance within the samples. That is, in most regions of the image, there are no objects present. Conversely, BCE Loss optimizes the model's predictions of changes and non-changes. Through the amalgamation of BCE Loss and Dice Loss, the model can attain the optimal solution of the objective function under foreground-background class imbalance conditions, thereby facilitating model convergence. In the following formula, y refers to the vector of true binary labels, $\hat{y}$ refers to the vector of predicted probabilities, N refers to the total number of samples, and i refers to the i-th training or inference sample. The formulations for BCE Loss and Dice Loss are presented as outlined below:

$$L_{BCE}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y})\right] \tag{19}$$

$$L_{Dice}(y, \hat{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} y_i^2} \tag{20}$$

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3502768

TABLE II
QUANTITATIVE ANALYSIS OF RSCD MODELS (PRECISION, RECALL, F1, IOU) ON THREE DATASETS. THE BEST VALUES ARE IN BOLD. ALL RESULTS ARE DESCRIBE AS PERCENTAGES (%)

.

Dataset	Params(M)	FLOPs(G)	FPS	LEVIR-CD				SYSU-CD				WHU-CD				CLCD			
				Pre	Rec	F1	IoU	Pre	Rec	F1	IoU	Pre	Rec	F1	IoU	Pre	Rec	F1	IoU
FC-Siam-Diff [33]	4.73	1.35	**169.89**	90.52	86.96	88.70	79.70	**91.47**	54.16	68.03	51.56	83.04	76.68	79.74	66.30	69.99	57.67	63.23	46.23
FC-Siam-Conc [33]	5.33	1.55	167.16	88.97	85.41	87.15	77.23	81.84	77.15	79.43	65.87	81.49	72.42	76.69	62.19	64.26	59.06	61.55	44.46
SNUNet [50]	12.03	54.82	45.74	90.06	87.20	88.61	79.55	74.01	81.15	77.41	63.15	84.70	66.80	74.69	59.60	62.42	65.60	63.97	47.03
DSIFN [25]	35.99	79.98	41.44	92.21	86.45	89.57	**88.11**	82.95	78.48	80.65	67.57	92.55	87.62	90.02	81.85	69.71	70.72	70.21	54.10
BIT [51]	3.01	16.88	66.50	89.91	88.63	89.26	80.61	80.65	73.96	77.16	62.81	80.97	84.95	82.91	70.81	55.93	64.29	59.82	42.67
TFI-GR-Net [52]	27.78	9.74	32.02	92.09	89.60	90.83	83.20	85.12	81.69	83.37	71.49	94.70	93.16	93.92	88.54	**82.70**	73.84	78.02	63.96
DMINet [53]	6.24	14.55	64.39	92.30	88.90	90.60	82.70	82.15	79.40	80.75	67.71	59.50	83.22	69.39	53.13	71.70	63.65	67.44	50.88
USSFC [42]	1.52	4.82	23.23	89.00	91.13	90.05	81.90	75.05	81.91	78.33	64.38	88.54	**94.39**	91.37	84.11	60.57	68.06	64.10	47.17
A2Net [41]	3.78	6.02	69.43	**93.17**	89.84	91.47	84.29	85.75	79.74	82.63	70.41	93.68	93.73	93.71	88.16	75.62	74.48	75.04	60.06
GAS-Net [54]	23.52	289.97	14.39	90.98	91.07	91.02	83.53	80.65	77.77	79.18	65.54	92.47	91.33	91.90	85.01	75.19	68.05	71.44	55.57
CLAFA [55]	5.02	16.60	46.86	92.45	89.72	91.06	83.59	80.37	**84.16**	82.23	69.82	74.30	87.23	80.25	67.02	75.43	72.76	74.07	58.82
CCL-Net [56]	5.23	8.69	37.99	88.86	**91.75**	90.29	82.29	83.01	73.47	77.95	63.87	79.07	89.53	83.98	72.38	62.99	60.71	61.83	44.75
EAFH-Net(ours)	4.81	2.93	78.00	92.36	91.49	**91.74**	84.74	85.49	83.10	**84.28**	**72.83**	**96.26**	93.91	**95.07**	**90.60**	80.16	**78.28**	**79.21**	**65.57**

Algorithm 1 Flowchart of EAFH-Net

Input:
$The\ set\ of\ image\ pairs\ T_1^n, T_2^n\ from\ the\ Nth\ batch\ along\ with\ their\ labels\ Y_L\ batch\ size\ B, maximum\ training\ iteration\ M, and\ learning\ rate\ lr$

output:
$four\ 256 \times 256\ change\ maps,\ learned\ under\ the\ supervision\ of\ Y_L.$

for $[1, M]$ **do:**

backbone extract feature :
$\{feat_1^{T1}, feat_2^{T1}, feat_3^{T1}, feat_4^{T1}, feat_5^{T1}\}$
$\{feat_1^{T2}, feat_2^{T2}, feat_3^{T2}, feat_4^{T2}, feat_5^{T2}\}$

After the steps in Eq.1-2 :
$\{D_2, D_3, D_4, D_5\}$

After the steps in Eq.3 :
$\{D_2', D_3', D_4', D_5'\}$

The upsampled change maps are obtained:
$C \in \{C_2, C_3, C_4, C_5\}$
$The\ loss\ L_{Bce}\ and\ L_{Dice}\ of\ calculatin\ C\ as\ Eq.19$
$-20,\ The\ total\ loss\ is\ calculated\ by\ summing\ the\ two\ losses\ and\ then\ backpropagating.$

end for

$$L_{total}(y, \hat{y}) = \sum_{i=2}^{5} [L_{Dice}(y_i, \hat{y}) + L_{BCE}(y_i, \hat{y})] \tag{21}$$

The Algorithm 1 provides an overview of the general workflow of EAFH-Net.

V. EXPERIMENTS

In this section, we commence by delineating the datasets and benchmark models for comparison utilized in the experiments. Following that, we expound on the detailed experimental procedures and evaluation metrics. Finally, we delve deeply into analyzing the experimental outcomes.

A. Datasets

In an endeavor to illuminate the broad-ranging adaptability inherent in our proposed EAFH-Net model, our study entailed rigorous experimentation employing three prevalent public datasets, as well as our newly proposed UAV-CD+ dataset collected from low-altitude UAVs. These experiments were divided into two groups.

In summary, in the comparative experiments, we compared three U-Net-based methods, three recent advanced CD methods, two transformer-based derivative methods, and four lightweight CD method.

The initial set of experiments was conducted to validate the efficacy of EAFH-Net using three widely-utilized public datasets, thus demonstrating its effectiveness. The subsequent set of experiments aimed at validating the efficacy of EAFH-Net on our proposed UAV-CD+ dataset, which comprises low-altitude UAV aerial images. Furthermore, we compared our dataset with existing baselines to confirm its effectiveness. Detailed information regarding the datasets is provided below:

1) Public Datasets: To illustrate the generalization and robustness of our proposed model, we conducted experiments using LEVIR-CD [18], SYSU-CD [19], WHU-CD [20] and CLCD [21]. LEVIR-CD, containing a large number of building changes, was evaluated to demonstrate effectiveness in real-world building CD scenarios; SYSU-CD, encompassing various changes such as buildings, vegetation, and water bodies, was utilized to validate the model's effectiveness in

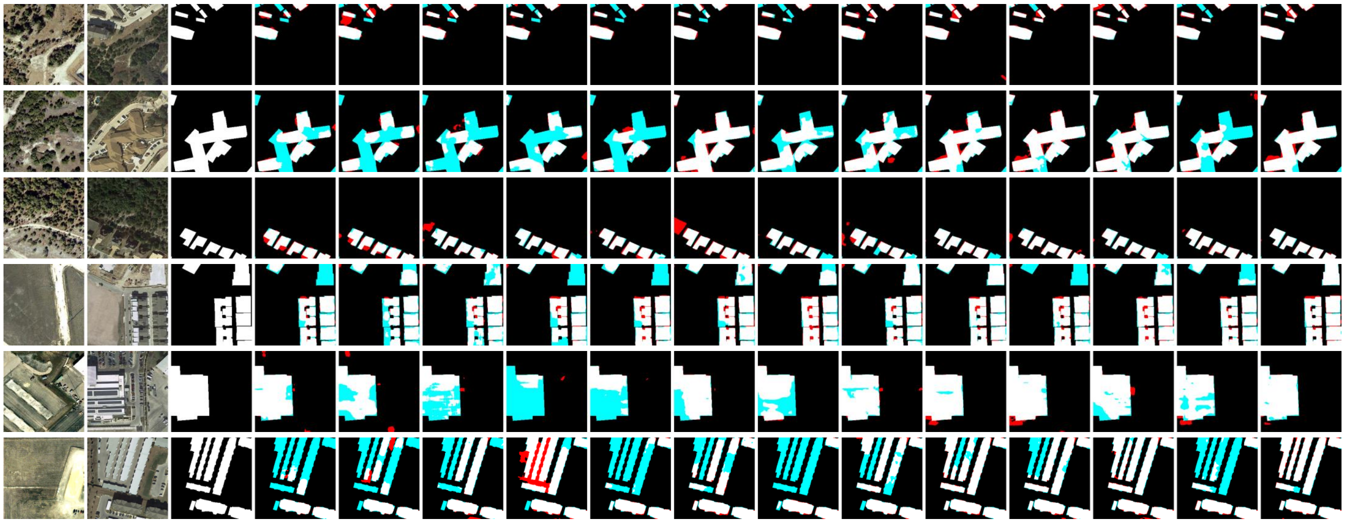


Fig. 6. Visual comparisons of the proposed method and the SOTA approaches on the LEVIR-CD dataset. (a) T_1 images. (b) T_2 images. (c) Ground truth. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) SNUNet. (g) DSIFN. (h) BIT. (i) TFI-GR-Net (j) DMINet (k) USSFC. (l) A2Net. (m) GAS-Net. (n) CLAFA. (o) CCL-Net (p) Ours. The colors of the images represent true positives (white), false positives (red), true negatives (black), and false negatives (blue).

complex change scenarios; CLCD, primarily focused on cropland changes but also including a small number of building, road, and lake changes, was employed to assess the model's sensitivity to cropland changes and minor complex scenarios. The LEVIR-CD dataset is characterized by its inclusion of 637 pairs of HR images, with each image exhibiting dimensions of 1024×1024 pixels. The images originate from diverse urban landscapes across Texas, USA, capturing temporal changes spanning from 2002 to 2018, and offer a spatial resolution of 0.5m/pixel. Samples are cropped into 256×256 pixel patches, meticulously partitioned into training, validation, and test subsets, maintaining a ratio of 7:1:2. The SYSU-CD dataset is composed of 20,000 pairs of orthographic aerial image patches, each with dimensions of 256×256 pixels and a spatial resolution of 0.5m. The patches are allocated across the training, validation, and test sets, maintaining an equitable distribution ratio of 6:2:2. CLCD comprises 600 pairs of agricultural land image samples, each boasting a pixel dimension of 512×512 pixels. The imagery was acquired within the confines of Guangdong Province, China, between the years 2017 and 2019, featuring a spatial resolution spanning from 0.5m to 2m. Change annotations cover various features including structures like buildings, roads, lakes, as well as barren expanses of soil, among others. The dataset is meticulously cropped into 256×256 pixel images, thereafter apportioned among the training, validation, and test sets following a ratio of 6:2:2. WHU-CD dataset, proposed by Wuhan University, is a high-resolution building CD dataset, which was collected in Christchurch, New Zealand. This dataset includes pairs of images taken in 2012 and 2016, with each image having dimensions of 32,507 × 15,354 pixels and a spatial resolution of 0.2 meters per pixel. In our experiment, the overlapping areas of these images were cropped into 256×256 patches. The dataset was then randomly split into training, validation, and test sets in an 8:1:1 ratio, resulting in 6096 images for training, and 762 images each for validation and testing.

2) UAV-CD+ Datasets: As explained in the third section, UAV-CD+ comprises 2002 pairs of low-altitude UAV images, each featuring a resolution of 1024×1024 pixels and an impressive spatial resolution of 0.06m. In the subsequent experiments, the images are cropped into several pairs of 256×256 images. Subsequently, the dataset is divided into training, validation, and test sets, with a proportion of 8:1:1. It encompasses a diverse range of change types captured in complex scenes from low-altitude aerial photography, as visualized in Fig. 2.

B. Comparative Methods

As illustrated in Fig. 6, we provide a comparison between several recently proposed CD networks and our proposed network with regards to model F1 score (F1), parameters (Params), and FLOPs. The models subjected to comparison: **FC-Siam-conc (FC-Conc) [33]**, **FC-Siam-diff (FC-Diff) [33]**, **SNU-Net [50]**, **DSIFN [25]**, **BIT [51]**, **TFI-GR-Net [52]**, **DMI-Net [53]**, **USSFC [42]**, **A2Net [41], GAS-Net [54], CLAFA [55] and CCL-Net [56].**

FC-Conc and **FC-Diff** are trailblazers within the realm of CD, and subsequent works are compared with them. **DSIFN**, **BIT**, **SNUNet**, and **TFI-GR-Net** are respectively the most classic transformer-based and CNN-based CD methods in recent years. **DMI-Net** and **GAS-Net** are the latest CD methods, while **USSFC**, **A2Net**, **CLAFA**, and **CCL-Net** are lightweight methods.

1) **FC-Conc**: Diverging from the conventional FC-EF, FC-Conc diverges by incorporating dual skip connections during the decoding process. It employs a Siamese encoder to capture features from bi-temporal images and routes these features from identical levels in the encoder to their corresponding decoder counterparts.

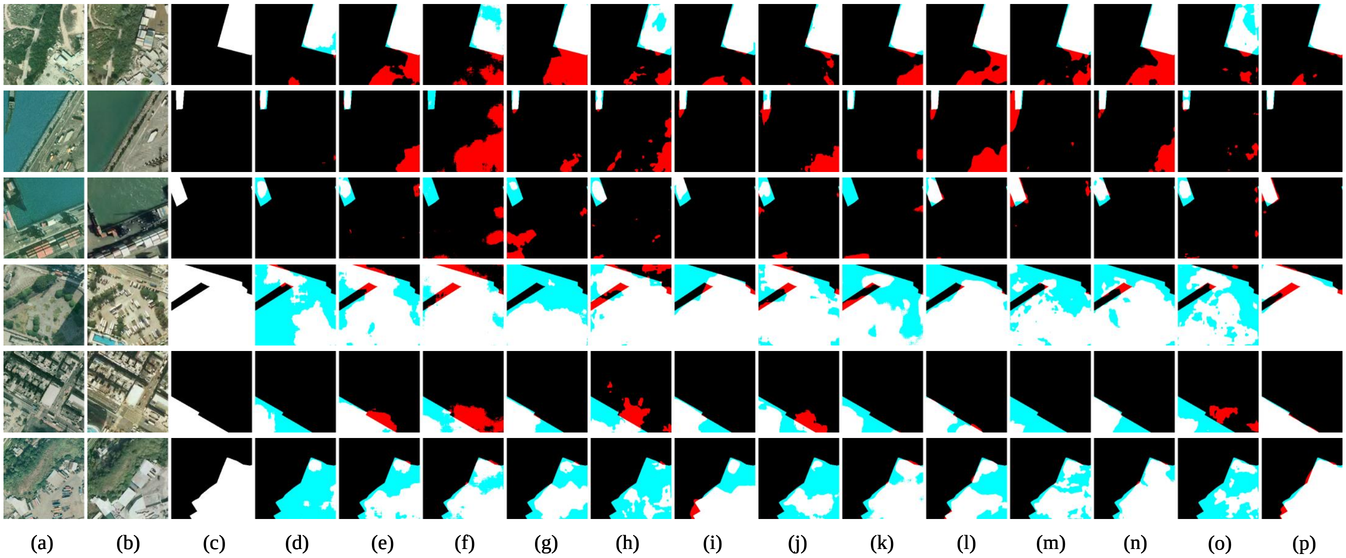


Fig. 7. Visual comparisons of the proposed method and the SOTA approaches on the SYSU-CD dataset. (a) T_1 images. (b) T_2 images. (c) Ground truth. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) SNUNet. (g) DSIFN. (h) BIT. (i) TFI-GR-Net (j) DMINet (k) USSFC. (l) A2Net. (m) GAS-Net. (n) CLAFA. (o) CCL-Net (p) Ours. The colors of the images represent true positives (white), false positives (red), true negatives (black), and false negatives (blue).

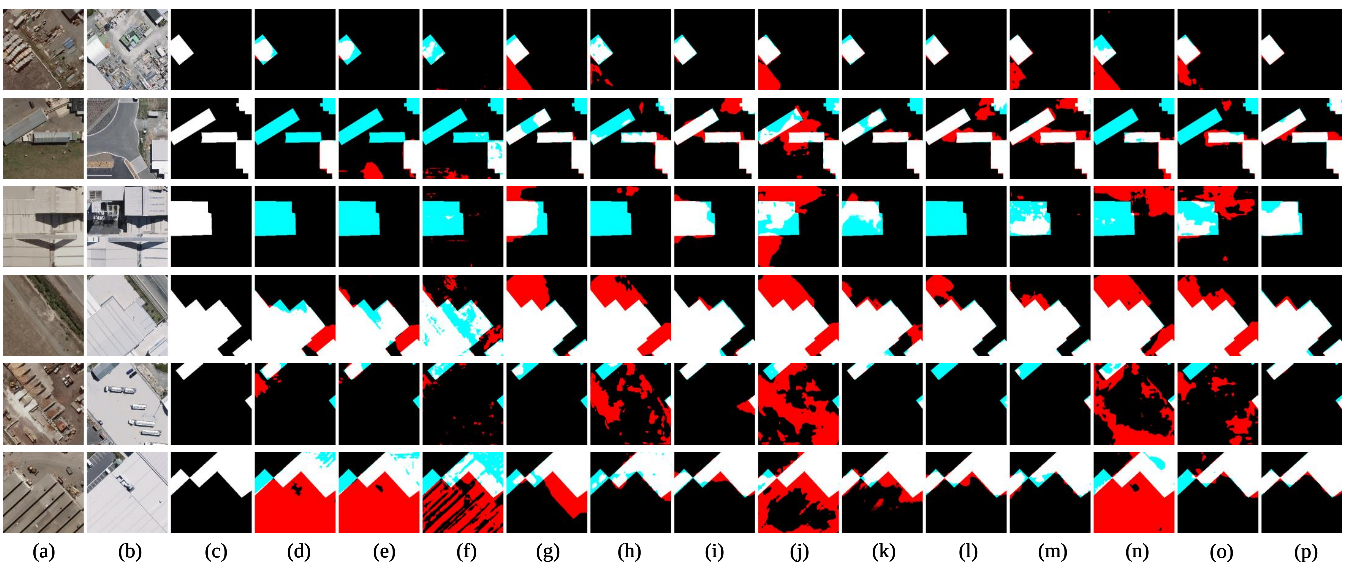


Fig. 8. Visual comparisons of the proposed method and the SOTA approaches on the WHU-CD dataset. (a) T_1 images. (b) T_2 images. (c) Ground truth. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) SNUNet. (g) DSIFN. (h) BIT. (i) TFI-GR-Net (j) DMINet (k) USSFC. (l) A2Net. (m) GAS-Net. (n) CLAFA. (o) CCL-Net (p) Ours. The colors of the images represent true positives (white), false positives (red), true negatives (black), and false negatives (blue).

2) **FC-Diff**: In contrast to FC-Conc, the FC-Diff module computes the absolute disparity of features extracted from corresponding tiers within the Siamese encoder, before propagating them to the decoder.
3) **SNU-Net**: By employing a Siamese network with dense connections for detection, the SNU-Net tackles the challenge of deep localization information loss by effectively transmitting information not only between the encoder and decoder but also within the decoder architecture.
4) **DSIFN**: The DSIFN represents a sophisticated deep supervised image fusion network crafted specifically for high-resolution bi-temporal RS image CD tasks. It utilizes a Siamese encoder to capture information from bi-temporal images and channels this information from identical levels in the encoder to their corresponding decoder counterparts. These deep features are subsequently propagated into a deep supervision difference discrimination network dedicated to CD.
5) **TFI-GR-Net**: The introduced CD network improves the interaction among bi-temporal features and captures temporal discrepancies across different feature levels utilizing a temporal feature interaction module. Additionally, it consolidates basic and advanced temporal difference representations via a guided refinement module, aiming to enrich the positional details of advanced-level features and eliminate background noise from basic-level features.

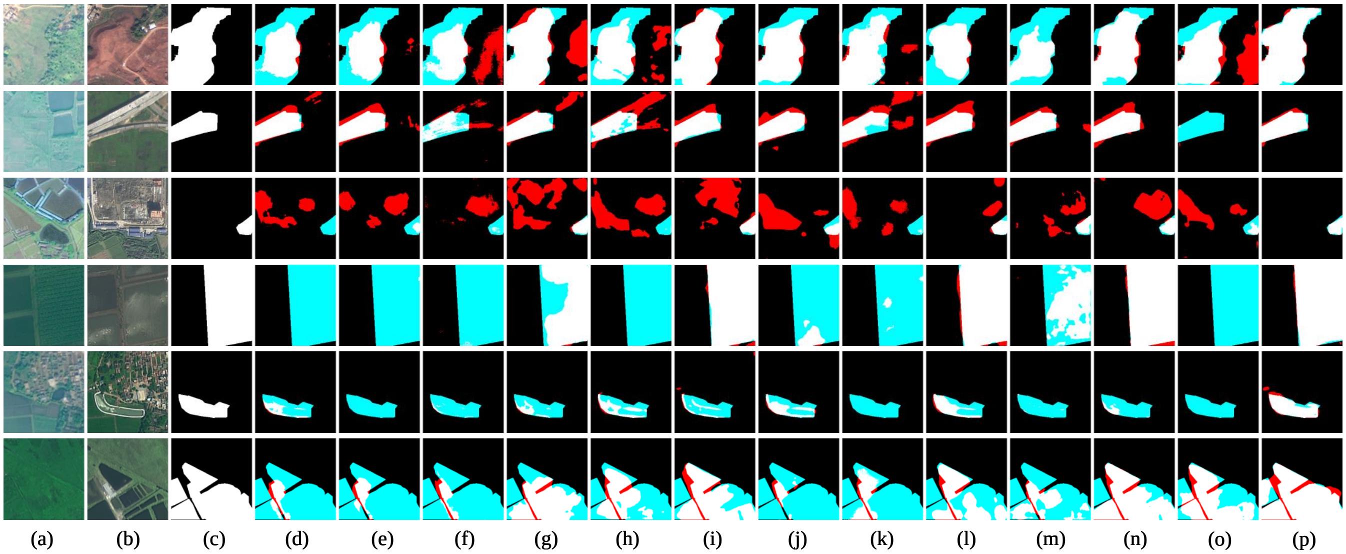


Fig. 9. Visual comparisons of the proposed method and the SOTA approaches on the CLCD dataset. (a) T_1 images. (b) T_2 images. (c) Ground truth. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) SNUNet. (g) DSIFN. (h) BIT. (i) TFI-GR-Net (j) DMINet (k) USSFC. (l) A2Net. (m) GAS-Net. (n) CLAFA. (o) CCL-Net (p) Ours. The colors of the images represent true positives (white), false positives (red), true negatives (black), and false negatives (blue).

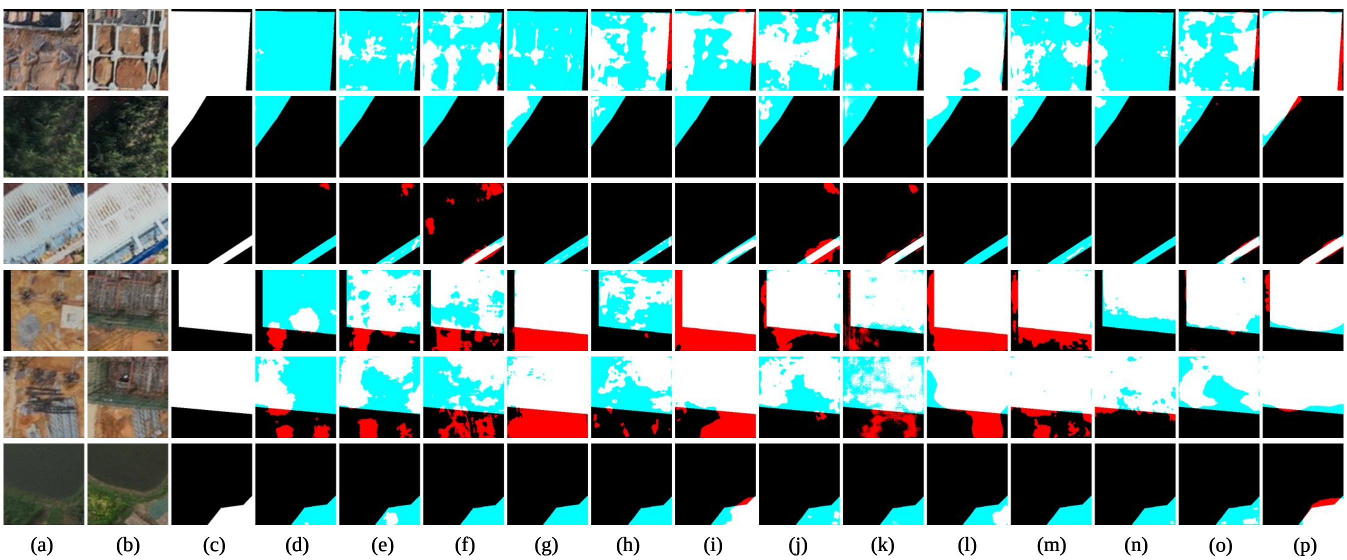


Fig. 10. Visual comparisons of the proposed method and the SOTA approaches on the UAV-CD+ dataset. (a) T_1 images. (b) T_2 images. (c) Ground truth. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) SNUNet. (g) DSIFN. (h) BIT. (i) TFI-GR-Net (j) DMINet (k) USSFC. (l) A2Net. (m) GAS-Net. (n) CLAFA. (o) CCL-Net (p) Ours. The colors of the images represent true positives (white), false positives (red), true negatives (black), and false negatives (blue).

6) **BIT**: Efficiently capturing spatio-temporal context within bi-temporal images is achieved through a bi-temporal image transformer. This model leverages a transformer encoder to succinctly model contextual information within a token-based spatio-temporal framework.
7) **DMI-Net**: This model embraces a hierarchical framework with multiple branches to efficiently and effectively generate representations of changes. The proposed method consolidates self-attention and cross-attention functionalities into a unified module, termed the cross-period joint attention block. This component orchestrates the overall feature distribution of each input, fostering the interaction between representations within individual layers and across different layers.
8) **USSFC**: Introducing a multi-scale decoupled CNN method, this approach adeptly captures features of changing objects at multiple scales, particularly when combined with the widely-used atrous spatial pyramid pooling (ASPP) technique.
9) **A2Net**: A novel lightweight network is proposed, which incorporates a neighbor aggregation module (NAM) to fuse nearby features in the backbone network, a progressive change identifying module (PCIM) to extract temporal difference information from bitemporal features, and a supervised attention module (SAM) to aggregate high-level and low-level features.
10) **GAS-Net**: Propose a network consisting of the global-attention module (GAM) and foreground-awareness mod-

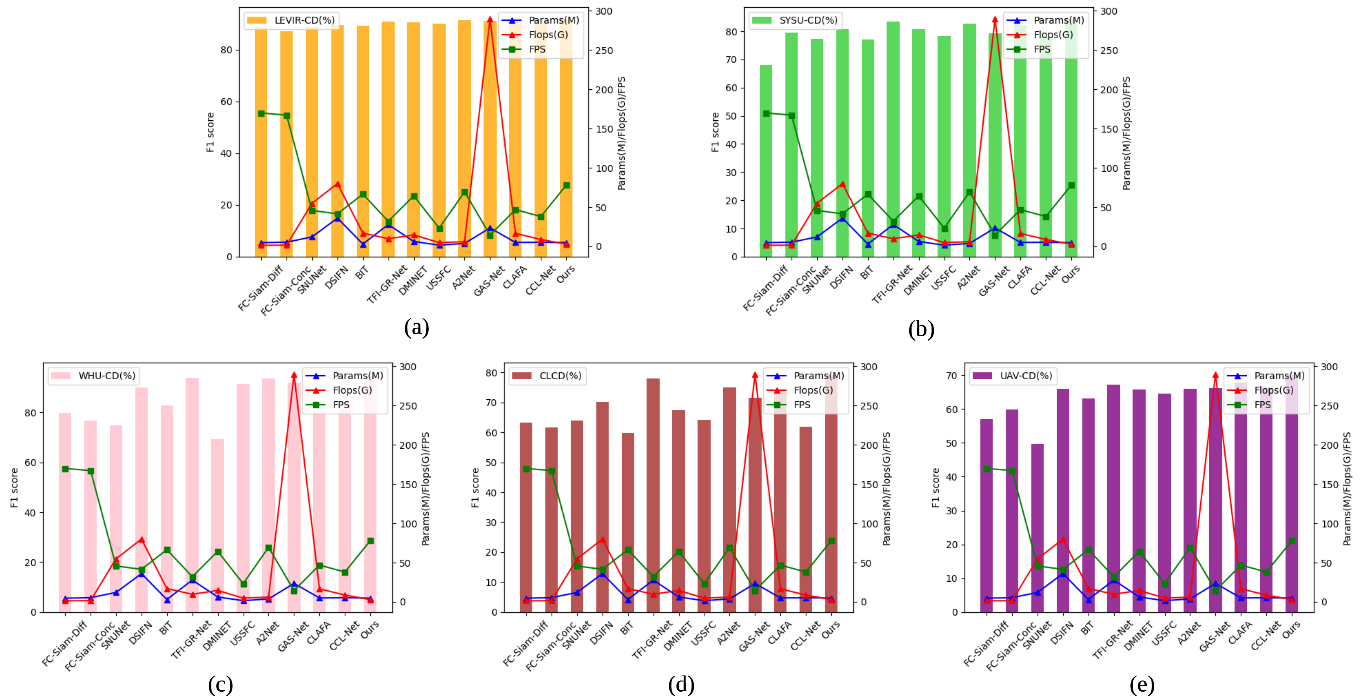


Fig. 11. Number of parameters and FLOPs of different methods and F1-scores of different methods on four datasets, (a) performance on LEVIR-CD; (b) performance on SYSU-CD; (c) performance on WHU-CD; (d) performance on CLCD; (e) performance on our proposed UAV-CD+ dataset.

ule (FAM), which both learn contextual relationships and enhance symbiotic relation learning between the scene and foreground.

11) **CLAFA**: Propose a network where the derived difference feature pyramid is progressively fused in an attend-then-filter manner, enabling high-level and adjacent lower-level features to constrain each other for stable feature fusion in identifying change regions. Additionally, an upsampling head is implemented with a stack of upsampling modules, each having a learnable branch that produces attentive residuals to refine statically upsampled results.

12) **CCL-Net**: Propose a cross-temporal context learning network that mines and fuses intratemporal and intertemporal long-range dependencies to exploit cross-temporal context information. Initially, a lightweight CNN extracts deep semantic features, followed by a cross-temporal fusion transformer (CFT) that locates changing objects by establishing long-range dependencies across bitemporal images.

All of the above networks have large Params and FLOPs, making them difficult to deploy for edge device CD tasks on UAVs. For example, even the smallest network, BIT, has Params and FLOPs of 3.01M and 8.48G, respectively. While achieving 89.26% detection accuracy on the LEVIR-CD dataset, its high computational cost and low accuracy make it unsuitable for high-precision and fast inference CD tasks on UAV edge devices. Therefore, there is a need to design a lightweight CD network with high accuracy, low FLOPS, and small Params.

TABLE III
QUANTITATIVE ANALYSIS OF RSCD MODELS (PRECISION, RECALL, F1, IOU) ON UAV-CD+. THE BEST VALUES ARE IN BOLD. ALL RESULTS ARE REPORTED IN PERCENTAGES (%)

Method(%)	Pre	Rec	F1	IoU
FC-Siam-Diff	71.93	47.20	57.00	39.86
FC-Siam-Conc	65.22	55.32	59.86	42.72
SNUNet	61.38	41.76	49.70	33.07
DSIFN	73.37	59.86	66.01	49.27
BIT	70.76	56.93	63.09	46.09
TFI-GR-Net	64.89	**69.80**	67.26	50.67
DMINet	67.89	63.78	65.77	48.99
USSFC	63.44	65.78	64.58	47.69
A2Net	74.32	59.24	65.93	49.17
GAS-Net	68.98	63.75	66.26	49.54
CLAFA	75.18	61.59	67.71	51.18
CCL-Net	74.95	59.43	66.30	49.58
EAFH-Net(Ours)	**77.13**	64.28	**70.12**	**53.99**

C. Implementation Details

The experimental setup utilized Ubuntu 22.04.3 LTS as the operating system, an RTX 2080Ti GPU, and PyTorch [58] version 1.8.0. Parameters for the compared methods were configured as closely as possible to the original literature. In cases of insufficient memory, the batch size was adjusted to 8. EAFH-Net was initialized with pre-trained weights from MobileNetV2, employing a batch size set to 32. A unified learning rate of 0.0005 was applied across all four datasets, employing the polynomial learning method:

$$(1-\frac{\text{current iteration}}{\text{max iteration}})^{\text{power}} \times \text{lr} \tag{22}$$

TABLE IV
ABLATION ANALYSIS OF RSCD MODELS (F1 AND IOU) ON FOUR DATASETS. THE BEST VALUES ARE IN BOLD. ALL RESULTS ARE DESCRIBE AS PERCENTAGES (%)

Variants	Params(M)	FLOPs(G)	LEVIR-CD		SYSU-CD		WHU-CD		CLCD		UAV-CD+	
			IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
EAFH-Net	4.81	2.93	84.74	91.74	**72.83**	**84.28**	**90.60**	**95.07**	**65.57**	**79.21**	53.99	70.12
(a) Baseline												
Baseline	2.25	0.88	66.13	79.61	54.77	70.78	68.94	81.62	47.30	64.22	37.69	54.74
(b) MFNFM												
MFNFM	4.05	1.77	83.57	91.05	71.72	83.53	88.14	93.70	63.21	77.46	46.12	63.13
All-Feature Fusion	5.34	3.99	84.44	91.57	70.90	82.97	88.44	93.87	62.33	76.80	49.13	65.89
Without concatenation	4.70	2.79	84.63	91.67	71.83	83.61	88.41	93.85	60.59	75.46	53.24	69.49
(c) SCFHM												
SCFHM	3.01	2.04	82.74	90.55	64.28	78.26	86.13	92.55	58.67	73.95	45.48	62.53
RSU7(1)	3.84	2.38	84.11	91.37	69.97	82.33	89.54	94.48	59.89	74.91	49.73	66.43
RSU7(2)	4.08	2.80	84.51	91.60	72.14	83.82	87.64	93.41	62.70	77.07	52.80	69.11
RSU4-4	4.22	2.93	**84.85**	**91.80**	72.78	84.24	88.99	94.18	64.41	78.36	51.26	67.78
RSU5-5	4.41	2.93	84.70	91.72	72.31	83.93	88.97	94.16	62.51	76.93	53.27	69.52
RSU6-6	4.6	2.93	84.59	91.66	72.37	83.97	88.33	93.81	63.88	77.96	49.06	65.82
(d)Difference backbone												
EfficientNet [59]	3.92	2.55	83.92	91.26	70.34	82.59	85.50	92.18	62.91	77.23	51.30	67.81
ShuffleNetV2 [60]	3.05	2.19	81.90	90.05	68.15	81.05	82.41	90.36	55.58	71.45	52.76	69.07
ResNet-18	13.96	6.93	83.92	91.26	72.03	83.74	88.14	93.69	62.91	77.23	53.48	69.69
ResNet-50	26.87	52.43	84.11	91.37	71.92	83.66	88.08	93.66	59.37	74.50	51.59	68.07
ResNet-101	44.06	90.54	83.94	91.27	71.60	83.45	84.25	91.45	64.64	78.52	**54.49**	**70.55**
(e)Loss												
Only BCE	-	-	84.13	91.38	71.94	83.68	89.44	94.43	65.14	78.89	53.51	69.72
Only Dice	-	-	84.36	91.52	70.69	82.83	89.87	94.67	60.78	75.61	49.83	66.52

The experiment proceeded for a maximum of 40,000 iterations, employing a power of 0.9 for the polynomial learning schedule. Adam optimizer [61] utilized a momentum of 0.9, a weight decay coefficient of 0.0001, and β_1 and β_2 parameters set to 0.9 and 0.99, respectively. Data augmentation strategies comprised random flipping, cropping, and temporal shuffling of input images.

To assess the efficacy of EAFH-Net, we employed five established metrics: precision (Pre), recall(Rec), F1-score (F1), and IoU. The specific formulas for computing these metrics as outlined below:

$$Pre = \frac{Tp}{Tp + Fp} \tag{23}$$

$$Rec = \frac{Tp}{Tp + Fn} \tag{24}$$

$$F1 = \frac{2Pre \cdot Rec}{Pre + Rec} \tag{25}$$

$$IoU = \frac{Tp}{Tp + Fp} \tag{26}$$

TP (true positives) represents correctly predicted positive samples, FP (false positives) represents incorrectly predicted positive samples, TN (true negatives) represents correctly predicted negative samples, and FN (false negatives) represents incorrectly predicted negative samples.

D. Performance Comparison on Public Datasets and UAV-CD+

1) The performance of all methods across the three public datasets with regards to parameters, FLOPs, Pre, Rec, F1, and IoU are presented in Table II. Our proposed model consistently surpasses other models while maintaining the lowest computational cost among the compared methods. For instance, on the LEVIR-CD, SYSU-CD, WHU-CD and CLCD datasets, EAFH-Net achieves F1 and IoU scores of 91.74% and 84.74%, 84.28% and 72.83%, 95.07% and 90.60%, 79.21% and 65.57%, respectively, with relatively low computational costs and high FPS. EAFH-Net surpasses USSFC on all four public datasets and other models with larger parameter sizes and higher computational costs, such as FC-Diff, FC-Conc, SNU-Net, DSIFN, BIT, TFI-GR-Net, DMI-Net, A2Net, GAS-Net, CLAFA and CCL-Net. Visual examinations of different approaches on the three public CD datasets are illustrated in Fig. 6, Fig. 7, Fig. 8, and Fig. 9.

2) Table III presents the effectiveness of each approach on the UAV-CD+ dataset in terms of Pre, Rec, F1, and IoU. EAFH-Net also achieves superior performance compared to other methods on the low-altitude UAV-captured dataset, with F1 and IoU of 70.12% and 53.99%, sequentially. Contrasted with USSFC, EAFH-Net achieves higher F1 and IoU. Moreover, in comparison to the approach with the highest F1, EAFH-Net achieves superior F1, suggesting that our

TABLE V
LOSS CONSTRAINT ABLATION ANALYSIS OF RSCD MODELS (F1 AND IOU) ON FOUR DATASETS. THE BEST VALUES ARE IN BOLD. ALL RESULTS ARE DESCRIBE AS PERCENTAGES (%)

Variants	LEVIR-CD		SYSU-CD		WHU-CD		CLCD		UAV-CD+	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
constraint(λ)										
0.1	84.30	91.48	72.50	**84.06**	89.30	94.39	64.43	78.37	51.74	68.19
0.2	84.50	91.60	72.15	83.82	89.64	94.54	64.69	78.56	53.17	69.42
0.3	84.58	91.64	72.38	83.98	89.75	94.60	65.21	78.94	52.74	69.06
0.4	84.27	91.47	72.31	83.93	89.32	94.36	65.26	**78.98**	53.58	69.77
0.5	84.34	91.51	72.19	83.85	89.62	94.52	65.25	78.97	53.94	**70.08**
0.6	84.51	91.60	72.03	83.74	89.51	94.46	64.64	78.52	52.98	69.26
0.7	84.48	91.59	71.58	83.43	90.09	94.78	64.98	78.77	51.84	68.28
0.8	84.38	91.53	72.26	83.89	89.86	94.66	63.86	77.94	52.19	68.59
0.9	84.74	**91.74**	71.87	83.63	90.25	**94.88**	64.67	78.54	51.10	67.63

approach achieves peak performance and accelerated computational speeds when processing high-altitude aerial imagery, RS satellite data, or low-altitude UAV-captured images. Fig. 10 provides a thorough visual comparison of all methodologies on the UAV-CD+ dataset.

E. Ablation Experiments

To thoroughly evaluate the performance of the EAFH-Net network modules, extensive ablation experiments were performed on three public datasets as well as the UAV-CD+ dataset, highlighting the influence of these modules. A comprehensive summary of the ablation experiment results across these four datasets is presented in Table IV.

1) Effectiveness of MFNFM: The efficacy of MFNFM stems from its capacity to amalgamate features extracted by neighboring sections of the backbone network, facilitating the fusion of intricate and elementary level features across layers. This integration enhances the extracted bi-temporal information and refines granularity details. To validate its effectiveness, we replaced MFNFM with a 1×1 convolutional block, ensuring dimension alignment without compromising the network's feature extraction capability. The outcomes delineated in Table IV unequivocally underscore the performance of MFNFM.

To discuss the effectiveness of MFNFM in integrating adjacent features, we input the last four layers of features extracted by the backbone into MFNFM. To ensure a fair comparison with the original MFNFM, we perform a 3x3 convolution on these four features for dimension alignment, followed by concatenation in pairs, as shown in Fig. 12. The ablation experiment of all-feature fusion is shown in Table IV. Although the parameters and FLOPs increase, the accuracy on the LEVIR dataset shows no significant change. However, the F1 score on the SYSD-CD dataset drops to 82.97%, on the WHU-CD dataset drops to 93.87%, on the CLCD dataset drops to 76.80%, and on the UAV-CD+ dataset drops to 65.89%. We speculate that the aggregation of features at all scales leads to over-smoothing, resulting in the loss of local details and important boundary information. Since the LEVIR dataset primarily contains building changes, the integration of new features has little impact. In contrast, the SYSD-CD and our proposed UAV-CD+ datasets involve a wider variety of changes, making them more prone to feature redundancy when new features are introduced. Additionally, the CLCD dataset, which contains many land changes, is more sensitive compared to the other datasets, as indicated by other ablation experiments. Therefore, the over-smoothing of features significantly affects the CLCD dataset. To validate the effectiveness of the Concatenation module we added, we conducted experiments after removing the Concatenation. The results are shown in Table IV.

2) Effectiveness of SCFHM: The SCFHM is pivotal in the harmonization and integration of spatial and channel bi-temporal features derived from the difference map. This integration accentuates the prominence of bi-temporal features within the feature space. Moreover, the insertion of an additional SCFHM between sophisticated and elementary levels features enables a more effective amalgamation of their spatial and channel bi-temporal characteristics during linking and fusion. This process culminates in the generation of more comprehensive features. Similarly, we used a 1×1 convolution to replace SCFHM, and the effectiveness of the two sets of SCFHM is shown in Table IV. In Table IV, RSU7(1) and RSU7(2) refer to the ablation test settings where the first and second sets of RSU blocks are removed from SCFHM, respectively. RSU4-4, RSU5-5, and RSU6-6 indicate the number of feature extraction repetitions in the two sets of RSU blocks. This setup is designed to demonstrate whether the depth of RSU blocks affects the final CD accuracy.

3) Discussion on the Convolutional Layer of SCFHM Module: The depth of layers in the convolutional layer of SCFHM directly influences its spatial and channel feature harmonization ability, as well as the size of parameters and computational costs. Consequently, we delve into the implications of the convolutional layer's depth within SCFHM on network performance. We conducted this experiment on three widely recognized public datasets, as shown in Table IV. We found that when the number of layers decreases, there is a small fluctuation in the increase or decrease of parameters and

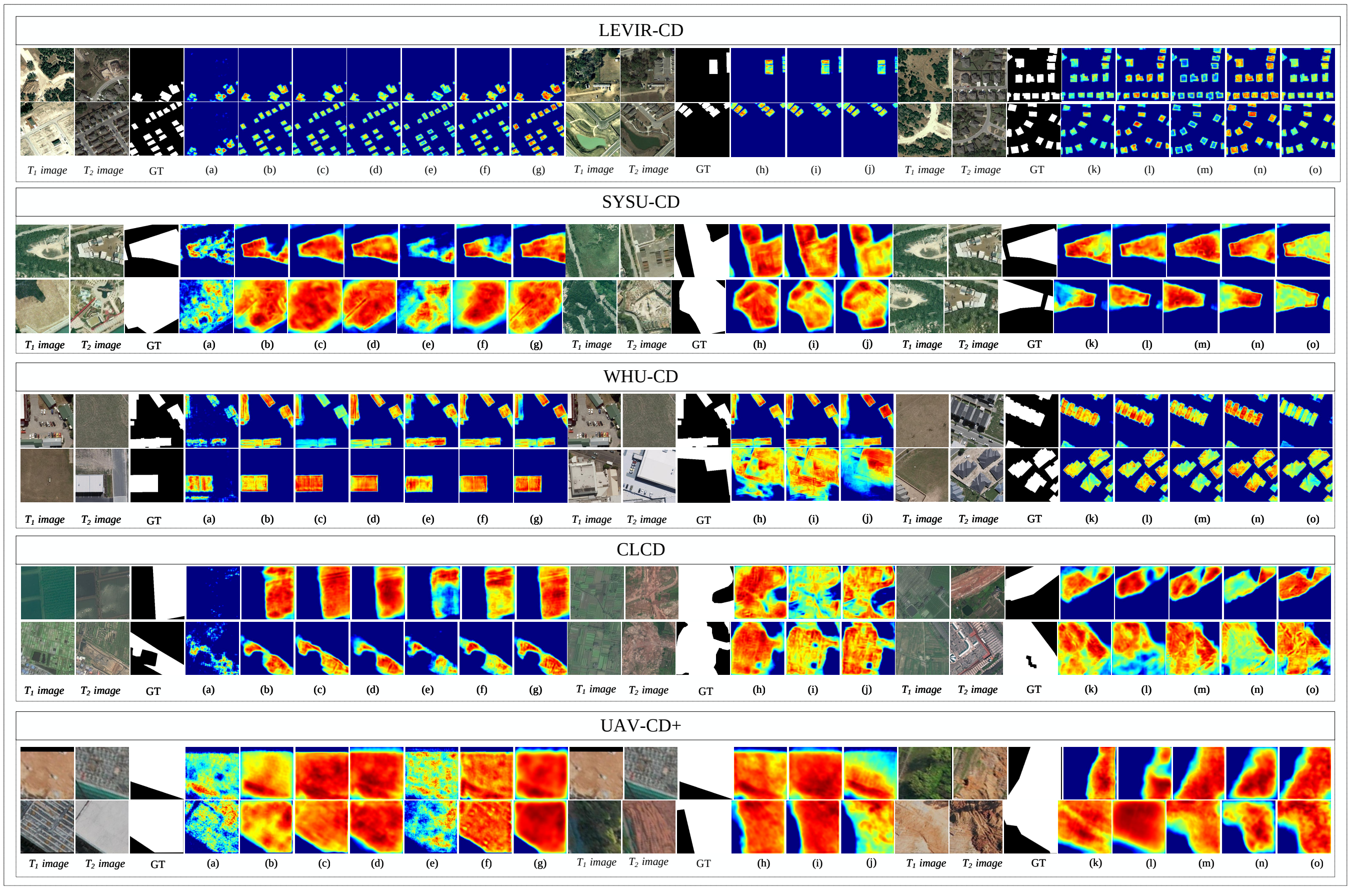


Fig. 12. Visual comparisons of the MFNFM and SCFHM and the ablation baseline modules in (a),(b),(c),(d),(e),(f),(g). Visual comparisons of different RSU layers in (h),(i),(j). Visual comparisons of different Backbone in (k),(l),(m),(n),(o). (a) Visual prediction of the baseline. (b) Visual prediction without MFNFM. (c) Visual prediction with All-Feature Fusion module. (d)Visual prediction without concatenation. (e) Visual prediction without SCFHM. (f) Visual prediction without the RSU7(1) in SCFHM. (g) Visual prediction without the RSU7(2) in SCFHM. (h) RSU4-4. (i) RSU5-5. (j) RSU6-6. (k)EfficientNet. (l)ShuffleNetV2. (m) Resnet-18. (n) Resnet-50. (o) Resnet-101. The above are all multi-scale fused feature heatmaps output from the decoder. It reflects the model's attention to areas of change.

computational costs, along with a slight decrease in accuracy. However, when both sets of SCFHM have 7 convolutional layers, their performance is more stable.

4) Discussion on the Impact of Different Backbones on the Network: We further validate different backbone networks fused with our proposed method, such as ResNet18, ResNet50 and ResNet101. The results from Table IV demonstrate that using MobileNetV2 as the backbone network, ialong with a slight decrease in accuracyts parameter count and computational cost are lower than ResNet18, ResNet50, and ResNet101, by 9.15M, 22.06M, and 39.25M, respectively, with corresponding reductions in FLOPs of 4G, 49.5G, and 90.54G. On LEVIR-CD, it exhibits F1 improvements of 0.48%, 0.37%, and 0.47%; on SYSU-CD, improvements of 0.54%, 0.62%, and 0.83%; on CLCD, improvements of 1.98%, 4.71%, and 0.69%; on UAV-CD+, improvements of 0.43% and 2.05%. However, when used as a backbone network on UAV-CD+, its F1 is lower by 0.43% compared to ResNet101. Nonetheless, as a lightweight network architecture, it still achieves commendable performance. This demonstrates that MobileNetV2 can attain excellent performance while ensuring network lightweightness. Since our proposed method emphasizes lightweight design, we incorporated ablation experiments focused on lightweight backbones. In these experiments, we replaced the EAFH-Net backbone with EfficientNet and ShuffleNetV2. The experimental results indicate that EfficientNet and ShuffleNetV2 exhibit weaker fitting capabilities for remote sensing bitemporal data compared to MobileNetV2.

5) Discussion on constraint conditions of hybrid loss function on the Network: To verify the effectiveness of the hybrid loss function, we conducted ablation experiments on cross-entropy loss and Dice loss separately. The implementation data is shown in Table IV. From the ablation experiments, it can be observed that the hybrid use of cross-entropy loss and Dice loss performs the best across the five datasets. Additionally, we added constraints to the Dice loss, with the constraints and experimental details provided in Table V.

VI. Conclusion

In this article, this article introduce EAFH-Net, a novel lightweight CD network, along with UAV-CD+, a benchmark dataset captured by low-altitude drones. UAV-CD+ significantly enhances existing CD datasets by addressing deficiencies in pixel resolution, spatial resolution, change types, and dataset volume for low-altitude captured data. EAFH-Net comprises an MFNFM module and two sets of SCFHM

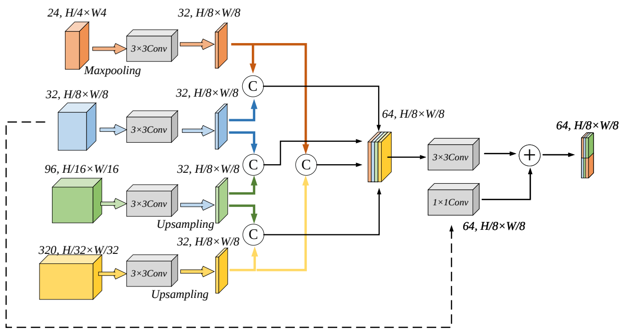


Fig. 13. The overview of MFNFM (All-Feature Fusion Version).

modules. MFNFM facilitates the fusion of features extracted by neighboring extraction of the backbone network, enabling cross-fusion of sophisticated and elementary levels features, thereby enhancing bi-temporal information extraction and refining granularity details. The first set of SCFHM harmonizes and integrates spatial and channel bi-temporal features of the difference map, accentuating the prominence of bi-temporal features in the feature space. The other set of SCFHM, inserted between sophisticated and elementary levels features, enhances the combination of their spatial and channel bi-temporal features during linking and fusion, resulting in more comprehensive features. Experimental results demonstrate that EAFH-Net achieves faster computational speed and superior performance on three public CD datasets and UAV-CD+, illustrating its adaptability to RS satellite images, high-altitude aerial images, and UAV-captured datasets, thereby showcasing the generality and robustness of our proposed network. In future endeavors, we will further optimize the dataset's change variety and the network's complexity and computational speed.

References

[1] H. Liu, M. Yang, J. Chen, J. Hou, and M. Deng, "Line-constrained shape feature for building change detection in vhr remote sensing imagery," *ISPRS International Journal of Geo-Information*, vol. 7, no. 10, p. 410, 2018.

[2] C. Marin, F. Bovolo, and L. Bruzzone, "Building change detection in multitemporal very high resolution sar images," *IEEE transactions on geoscience and remote sensing*, vol. 53, no. 5, pp. 2664–2682, 2014.

[3] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 1, pp. 125–138, 2015.

[4] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5407–5423, 2017.

[5] Y. Zhou, "Research on forest resource change detection based on decision tree algorithm," in *2022 International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*. IEEE, 2022, pp. 363–367.

[6] C.-F. Chen, N.-T. Son, N.-B. Chang, C.-R. Chen, L.-Y. Chang, M. Valdez, G. Centeno, C. A. Thompson, and J. L. Aceituno, "Multi-decadal mangrove forest change detection and prediction in honduras, central america, with landsat imagery and a markov chain model," *Remote Sensing*, vol. 5, no. 12, pp. 6408–6426, 2013.

[7] Y. Zheng, X. Zhang, B. Hou, and G. Liu, "Using combined difference image and k-means clustering for sar image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 3, pp. 691–695, 2013.

[8] G. H. Dunteman, *Principal components analysis*. Sage, 1989, vol. 69.

[9] 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 Geoscience and Remote Sensing Letters*, vol. 13, no. 5, pp. 681–685, 2016.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[13] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[17] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8440–8449.

[18] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[19] 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 transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021.

[20] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on geoscience and remote sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[21] M. Liu, Z. Chai, H. Deng, and R. Liu, "A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4297–4306, 2022.

[22] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. Ieee, 2018, pp. 2115–2118.

[23] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 565–571, 2018.

[24] 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 Sensing*, vol. 13, no. 24, p. 5094, 2021.

[25] 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 Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[26] Q. Wang, Z. Yuan, Q. Du, and X. Li, "Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2018.

[27] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[28] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 193–15 202.

[29] H. Lee, K. Lee, J. H. Kim, Y. Na, J. Park, J. P. Choi, and J. Y. Hwang, "Local similarity siamese network for urban land change detection on remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4139–4149, 2021.

[30] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[31] K. Zhang, X. Zhao, F. Zhang, L. Ding, J. Sun, and L. Bruzzone, "Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[33] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[34] R. Barkur, D. Suresh, S. Lal, C. S. Reddy, P. Diwakar *et al.*, "Rscdnet: A robust deep learning architecture for change detection from bi-temporal high resolution remote sensing images," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 2, pp. 537–551, 2022.

[35] C. Zhao, Y. Tang, S. Feng, Y. Fan, W. Li, R. Tao, and L. Zhang, "High-resolution remote sensing bitemporal image change detection based on feature interaction and multitask learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[36] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9774–9788, 2023.

[37] Z. Li, S. Cao, J. Deng, F. Wu, R. Wang, J. Luo, and Z. Peng, "Stade-cdnet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[38] Y. Deng, J. Chen, S. Yi, A. Yue, Y. Meng, J. Chen, and Y. Zhang, "Feature-guided multitask change detection network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9667–9679, 2022.

[39] J. Ding, X. Li, S. Xiang, and S. Chen, "Multi-level features fused and change information enhanced neural network for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[40] L. Wang, L. Wang, Q. Wang, and L. Bruzzone, "Rscnet: A residual self-calibrated network for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[41] 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[42] 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[43] J. Li, S. Li, and F. Wang, "Lcdnet: Lightweight change detection network with dual attention guidance and multiscale feature fusion for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[44] X. Jiang, S. Zhang, J. Gan, J. Wei, and Q. Luo, "Lfhnet: Lightweight full-scale hybrid network for remote sensing change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[45] M. Wang, B. Zhu, J. Zhang, J. Fan, and Y. Ye, "A lightweight change detection network based on feature interleaved fusion and bi-stage decoding," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[46] Y. Dai, L. Shen, Y. Wang, S. Liu, and Z. Li, "A lightweight dual-branch network for building change detection in remote sensing images integrating cross-scale coupling and boundary-constraint," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[47] C. Serief, Y. Ghelamallah, and Y. Bentoutou, "Deep-learning-based system for change detection onboard earth observation small satellites," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 8115–8124, 2023.

[48] Z. Ying, Z. Tan, W. Li, J. Zhou, Z. Liang, and Y. Zhai, "Uav-bcd: A uav building change detection dataset," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 1012–1015.

[49] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.

[50] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[51] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[52] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

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

[54] Z. Song, X. Wei, X. Kang, S. Li, and J. Liu, "Toward efficient remote sensing image change detection via cross-temporal context learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–10, 2023.

[55] R. Zhang, H. Zhang, X. Ning, X. Huang, J. Wang, and W. Cui, "Global-aware siamese network for change detection on remote sensing images," *ISPRS journal of photogrammetry and remote sensing*, vol. 199, pp. 61–72, 2023.

[56] G. Wang, G. Cheng, P. Zhou, and J. Han, "Cross-level attentive feature aggregation for change detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[57] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[59] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[60] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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 Universita 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, Self Supervise Learning.

Jiahao Pan (Student Member, IEEE) received the B.S. degree from Wuyi University, Jiangmen, China, in 2022. He is pursuing the master's degree with the College of Electronics and Information Engineering, Wuyi University, Jiangmen, China. His research interests include computer vision, change detection and object detection.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3502768

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.

Tingfeng Xian (Student Member, IEEE) received the B.S. degree from Wuyi University, Jiangmen, China, in 2022. He is pursuing the master's degree with the College of Electronics and Information Engineering, Wuyi University, Jiangmen, China. His research interests include semantic segmentation, change detection and pattern recognition.

Ying Xu received her Ph.D. degree in automation from the South China University of Technology, Guangdong, China, in 2013. She is currently an Associate Professor with the School of Department of Intelligence Manufacturing, Wuyi University, Jiangmen, Guangdong, China.

Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in Industrial and Information Engineering from the Universita degli Studi della Campania Luigi Vanvitelli, Italy, in 2019. From 2019 to 2022, he was a post-doctoral researcher at the Universita degli Studi di Padova, Italy. He is currently an Assistant Professor at the Department of Computer Science of the Universita degli Studi di Milano, Italy. He is Co-chair of the Intelligent Measurement Systems Technical Committee (TC-22) of the IEEE Instrumentation and Measurement Society since 2023. His research activities are focused on theoretical, methodological, and applied aspects of computational intelligence for signal and image processing.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Universita degli Studi di Milan, Crema, Italy, in 2014. He has been a postdoctoral Research Fellow in computer science with the Universita degli Studi di Milan since 2014. 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 current research interests include 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.

Vincenzo Piuri (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer engineering from Politecnico di Milan, Milan, Italy, in 1984 and 1988, respectively. He was the Department Chair with the University of Milan, Milan, from 2007 to 2012, where he has been a Full Professor since 2000. He was an Associate Professor with Politecnico di Milan from 1992 to 2000, a Visiting Professor with The University of Texas at Austin, Austin, TX, USA from 1996 to 1999, and a Visiting Researcher with George Mason University, Fairfax, VA, USA, from 2012 to 2016. He founded a startup company, Sensuresrl, Bergamo, Italy, in the area of intelligent systems for industrial applications (leading it from 2007 to 2010) and was active in industrial research projects with several companies. His main research and industrial application interests are intelligent systems, computational intelligence, pattern analysis and recognition, machine learning, signal and image processing, biometrics, intelligent measurement systems, industrial applications, distributed processing systems, Internet-of-Things, cloud computing, fault tolerance, application-specifi digital processing architectures, and arithmetic architectures. Dr. Piuri is an ACM Fellow.

Fabio Scotti (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003. He was an Assistant Professor with the Department of Information Technologies, Universita degli Studi di Milan, Milan, from 2002 to 2015, where he was an Associate Professor with the Department of Computer Science from 2015 to 2020. He has been a Full Professor with the Universita degli Studi di Milan since 2020. Original results have been published in over 150 papers in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, 3-D reconstruction, industrial applications, intelligent measurement systems, and high-level system design. Prof. Scotti is an Associate Editor of the IEEE Transactions on Human–Machine Systems and the IEEE Open Journal of Signal Processing. He is serving as a Book Editor (Area Editor, section Less-Constrained Biometrics) of the Encyclopedia of Cryptography, Security, and Privacy (3rd Edition, Springer). He has been an Associate Editor of the IEEE Transactions on Information Forensics and Security, Soft Computing (Springer) and a Guest Coeditor of the IEEE Transactions on Instrumentation and Measurement.

C. L. Philip Chen (Fellow, IEEE) received the M.S. degree in electrical engineering from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988. He was a tenured Professor, the Department Head, and an Associate Dean of two different universities in the U.S. for 23 years. He is currently the Head of the School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China. His current research interests include systems, cybernetics, and computational intelligence. Dr. Chen received the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University. He has been the Editor-in-Chief of the IEEE TRANSACTION ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS since 2014 and an associate editor of several IEEE TRANSACTIONS. He was the Chair of TC 9.1 Economic and Business Systems of International Federation of Automatic Control from 2015 to 2017 and also a Program Evaluator of the Accreditation Board of Engineering and Technology Education of the U.S. for computer engineering, electrical engineering, and software engineering programs. He was the IEEE SMC Society President from 2012 to 2013 and a Vice President of Chinese Association of Automation (CAA). He is a Fellow of AAAS, IAPR, CAA, and HKIE. Dr. Piuri is an ACM Fellow.