FDA-FFNet: A Feature-Distance Attention-Based Change Detection Network for Remote Sensing Image

Wenguang Peng, Wenzhong Shi, Min Zhang, and Lukang Wang

***Abstract*—Convolutional neural networks have demonstrated remarkable capability in extracting deep semantic features from images, leading to significant advancements in various image processing tasks. This success has also opened up new possibilities for change detection (CD) in remote sensing applications. But unlike the conventional image recognition tasks, the performance of AI models in CD heavily relies on the method used to fuse the features from two different phases of the image. The existing deep-learning-based methods for CD typically fuse features of bitemporal images using difference or concatenation techniques. However, these approaches often fail tails to prioritize potential change areas adequately and neglect the rich contextual information essential for discerning subtle changes, potentially leading to slower convergence speed and reduced accuracy. To tackle this challenge, we propose a novel feature fusion approach called feature-difference attention-based feature fusion CD network. This method aims to enhance feature fusion by incorporating a feature-difference attention-based feature fusion module, enabling a more focused analysis of change areas. Additionally, a deep-supervised attention module is implemented to leverage the deep surveillance module for cascading refinement of change areas. Furthermore, an atrous spatial pyramid pooling fast is employed to efficiently acquire multiscale object information. The proposed method is evaluated on two publicly available datasets, namely the WHU-CD and LEVIR-CD datasets. Compared with the state-of-the-art CD methods, the proposed method outperforms in all metrics, with an intersection over union of 92.49% and 85.56%, respectively.**

***Index Terms*—Attention-based, change detection (CD), deep learning, deep supervision, multiscale feature.**

Manuscript received 18 October 2023; revised 30 November 2023; accepted 14 December 2023. Date of publication 19 December 2023; date of current version 3 January 2024. This work was supported in part by Otto Poon Charitable Foundation Smart Cities Research Institute, the Hong Kong Polytechnic University (Work Program: CD03), in part by the Beijing Key Laboratory of Urban Spatial Information Engineering under Grant 2020101, and in part by The Hong Kong Polytechnic University under Grant 1-ZVN6, Grant ZVU1, and Grant U-ZECR. *(Corresponding author: Wenguang Peng.)*

Wenguang Peng is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: wenguangpeng@whu.edu.cn).

Wenzhong Shi and Min Zhang are with the Smart Cities Research Institute, Hong Kong Polytechnic University, Hong Kong, and also with the Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong (e-mail: john.wz.shi@polyu.edu.hk; lsgi-min.zhang@polyu.edu.hk).

Lukang Wang is with the School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China (e-mail: wanglukang@cumt.edu.cn).

The codes are available at https://github.com/pwg111/FDAFFNet.

Digital Object Identifier 10.1109/JSTARS.2023.3344633

I. Introduction

CHANGE detection (CD) is the process of comparing multiple raster datasets captured at different time phases to identify and quantify the extent of changes occurring in a given area. These changes can vary from anthropogenic landscape modifications and sudden natural disasters to long-term climate or environmental shifts. The investigation of CD using remote sensing images has been a prominent research area within the field of remote sensing for several years. It has been applied to diverse fields, such as land use [1], [2], urban expansion [3], [4], urban planning [2], geologic disaster monitoring (including floods [5] and landslides [6], [7]), ecological environmental protection, wetland monitoring [8], and forest protection [9], [10]. With the rapid development of optical sensors, automatic CD technology is gaining increasing attention for its potential to reduce human and material costs in acquisition and analysis process.

Traditional CD methods can be roughly categorized into three categories: direct comparison, postclassification comparison, and direct classification. Among these, the direct comparison method is the most commonly employed approach. It operates on the assumption that the spectral features of unchanged areas in two images should remain consistent or similar over time. By conducting various operations and transformations on aligned pixel values from the two bitemporal remote sensing images, areas of change can be identified. Direct comparison methods encompass various techniques, such as the pixel difference method [11], image regression method [12], and waveband substitution method. On the other hand, the postclassification comparison method involves an initial classification of the bitemporal remote sensing images, followed by a comparison of the classification results to identify changes. The accuracy of this method relies on both the classification method's accuracy and the consistency of the classification standard used. Therefore, it is crucial to maintain the consistency between the classification method and the classification standard. Another category, the direct classification method, combines the concepts of the image direct comparison method and the postclassification result comparison method. Common methods within this category include change vector analysis (CVA) [13], support vector machine (SVM) [14], tasseled cap transformation (KT) [15], and multitemporal phase combination postclassification methods. However, these traditional approaches have limitations in capturing spatial

contextual information and complex visual features. They often rely on accessing dimensional features after RGB and linear transformations, which hinders the extraction of deep-level change features. Additionally, the traditional CD methods typically require experts with professional knowledge and experience to perform feature extraction and selection, resulting in limited efficiency and robustness.

With the continuous advancements in computer vision and deep learning, convolutional neural networks (CNNs) have proven to be highly effective in capturing intricate details and complex texture features in images. They have achieved remarkable success in various imaging tasks, such as object detection and segmentation. CNN-based models typically follow an encoder–decoder architecture. The encoder, which includes popular architectures, such as VGG [16], ResNet [17], DarkNet [18], MobileNet [19], [46], and others, extracts both deep and shallow hierarchical information from the input image. The final recognition result is obtained through a dedicated decoder. In the context of image segmentation, the decoder usually employs a stage-by-stage upsampling technique to generate a probability-based mask map. In object detection, the decoder typically takes the form of an anchor-free or anchor-based detector head, generating bounding boxes, various confidence scores, and confidence levels of objects. Training CNNs usually requires a large amount of data. Various augmentation methods are available to increase the dataset size, enhance recognition accuracy, and improve the robustness of CNNs. These methods include image blurring, color space transformations, image flipping, multisample fusion, and Mosaica methods [20]. CNN models are highly sensitive to both the quantity and quality of the data. As remote sensing enters the era of big data, CNNs have proven effective in remote sensing image recognition tasks, such as building extraction, scene classification, texture evaluation, and other Earth science applications. Deep-learning methods have demonstrated their effectiveness in these tasks as well. Over recent years, deep-learning methods have gradually become the predominant focus of research in remote sensing imagery due to their higher accuracy, faster processing speed, and greater robustness compared with traditional methods.

In the domain of CD, there is a fundamental difference compared with general image research: CD involves the input of two images, whereas other image domains typically deal with a single image or video. CNNs have been widely used in the CD domain due to their weight-sharing property. Among the CNN architectures, Siamese networks [21] have emerged as the preferred choice for CD in remote sensing images. They consist of twin encoders and a single decoder, allowing for weight sharing between the encoders to constrain the feature learning. These networks employ various designed feature fusion methods to integrate the features from different time phases. Ultimately, the fused features are fed into the decoder to obtain the final change classification results. However, the existing CD models tend to prioritize to study spatiotemporal changes in images rather than expending excessive energy on feature extraction of diachronic images. Moreover, objects in remote sensing images often exhibit significant variations in scale, requiring a model with a high level of capability for extracting features at multiple scales. An efficient and effective model that specifically focuses on capturing differences becomes necessary. Additionally, the feature interaction before capturing disparity is significantly motivated and supported by the uncertainty in the distribution of changes in bistatic images. It is also important to suppress task-irrelevant disturbances, such as seasonal turnover and building remodeling. Furthermore, when addressing the imbalance between target and background categories, particularly in the real-world production processes, it is crucial to consider that many regions of the world have undergone extensive urbanization and development. Consequently, there are relatively few interannual changes in urban landscapes. Therefore, the feature fusion process should prioritize change regions to enhance the model's convergence speed and improve recognition accuracy.

To address the problem above, we propose a feature-difference attention-based feature fusion CD network (FDA-FFNet), and we introduce a novel feature fusion module and a deep-supervised attention mechanism (DSAM). The feature-difference attention-based feature fusion module (FDA-FFM) enables a more focused approach toward change areas. Additionally, we incorporate a DSAM to enhance the utilization of the deep surveillance module, allowing for cascading refinement of change areas. Our main contributions are given as follows.

1) A novel CD model is proposed, termed FDA-FFNet, which focuses on identifying potential changes and progressively supervising multilevel features using DSAM for more refined CD.
2) An FDA-FFM is designed to guide the fusion of features from both deep and shallow levels in the bitemporal images, and a DSAM is implemented to effectively utilize the deep surveillance attention module for cascading refinement of change areas. Spatial pyramid pooling fast (SPPF) is introduced into the model to enhance its capability of extracting multiscale target features.
3) Comparative experiments and ablation experiments are conducted on the WHU-CD and LEVIR-CD datasets. The proposed method performs well on all metrics, achieving an intersection over union (IoU) of 92.49% and 85.56%, respectively. Furthermore, side-by-side ablation experiments are carried out to validate the effectiveness of the proposed feature fusion module.

II. Related Work

A. Traditional CD Methods

In the early stages, the technology of deep learning was not mature enough. Additionally, due to the limitations in sensor technology, early remote sensing images only had low or medium spatial resolution. This lack of high-resolution data hindered the creation of large and high-quality dataset. As a result, traditional methods, mainly algebra-based methods [10], [11], [12], [13], [14], dominated the field. Among them, RSCD techniques were proposed specifically for detecting change in these types of images. Image algebra-based methods commonly utilize techniques, such as image differencing, image regression, CVA, SVM, and KT. These methods typically require the selection of an appropriate threshold to identify regions of

change. Moreover, expert intervention is necessary for feature extraction and selection, relying on their professional knowledge and experience.

B. Deep-Learning-Based CD Methods

The full convolutional network (FCN) [22] is an early deep-learning framework proposed for segmentation. It is considered a pioneering work in the field of semantic segmentation, as it was one of the first applications of deep learning in this area. In FCN, the fully connected layer of the traditional CNN is replaced by a convolutional layer. This modification enables the network to generate a heatmap instead of classifying categories. Additionally, to address the issue of image size reduction caused by convolution and pooling, an upsampling method is employed to restore the original image size. CD is often approached as a pixel segmentation task. FCN is widely used for CD tasks and is commonly categorized into single-stream and dual-stream networks.

Single-stream networks, which typically take a concatenation or difference of two-phase images as input, can be utilized in conjunction with standard semantic segmentation networks. In research, it is common practice to customize these networks to incorporate specific features. Popular networks used for semantic segmentation include UNet [23], among others. The simplest approach to customization involves modifying the number of input channels in the initial convolutional layer of the network to align with the number of resulting channels after operations, such as concatenating or differencing two-phase images. Daudt et al. [21] first proposed a single branch input fully CNN structure based on UNet; he directly connects the image pairs in channel dimension as inputs and completes the CD by using the underlying UNet network. Zhou et al. [24] similarly channel-connected inputs the aligned images into the UNet++ to extract feature maps with high spatial accuracy, and finally fused the multilateral outputs to obtain the final CD prediction using superposition; Zhang et al. [25] first establish the difference pyramid of the input dual-time-phase image, and then use the difference pyramid as the input of UNet++, and also use the strategy of fusing different scales of multiple inputs to complete the CD.

Dual-stream networks leverage the weight-sharing property of CNNs. These networks typically consist of an encoder with two feature extraction networks that share weights, along with a specific decoder to produce the change result. This structure is widely adopted as the mainstream network structure for CD. Daudt et al. [21] were the first to propose two fully convolutional two-stream networks in which bidimensional temporal features are extracted in a Siamese manner. Subsequently, pairs of features are fused in various ways before being passed to the decoder to reconstruct the change map. Zhang et al. [26] used a twin VGG as the backbone network for feature extraction from two-phase images. These features were then fed into a feature fusion subnetwork that utilized a combination of channel and spatial attention, along with a deeply supervised strategy to progressively refine the change maps. Shi et al. [27] used ResNet as the backbone network. They computed feature distances based on the output of the first two layers of ResNet and supervised it using Dice loss. They obtained the final change map through the employment of the CBAM module. Zhang et al. [28] proposed a Siamese network with a hierarchical fusion strategy. Bitemporal features were hierarchically fused with connectivity options. Fang et al. [29] proposed a CD network similar to UNet++. Although the weight-sharing property of the two-stream network feature extraction network is utilized, the performance of the features in unchanged areas, across different feature channels, can still vary considerably. The fusion method for the two-phase image features is a crucial factor that affects the accuracy of the image model.

To address this, many researchers have explored difference fusion methods. Feng et al. [30] divided the encoder into two segments and incorporated the JointAtt module in the first segment. This module utilized the two-phase image features output from the first encoder as input and generated two attentions to be used in the second encoder. This approach allows the model to pay more attention to potential change regions; Fang et al. [31] proposed a feature interaction mechanism that includes spatial feature interaction and channel feature interaction. These interactions are used to fuse the two-phase features. Li et al. [32] used MobileNetv2 as an encoder and designed modules, such as NAM and PCIM, to enhance the model's feature extraction capability from a lightweight perspective, resulting in improved performance; Zhu et al. [33] used an encoder–decoder Siamese network to extract features from bitemporal images and select balanced training samples through a global hierarchical sampling mechanism; in addition, the method also incorporates a binary change mask in the decoder to attenuate the influence of the background of the unchanged region on the foreground of the changed region, which further improved the detection accuracy.

III. Methodology

A. Overview

The proposed model follows an encoder–decoder structure, as illustrated in Fig. 1. The encoder is composed of a weight-sharing twin ResNet, SPPF, and FDA-FFM. On the other hand, the decoder, known as the change analysis network, utilizes stepwise upsampling convolutional networks and a DSAM to obtain the CD results.

In practice, the first four convolutions of ResNet [17] are utilized as the backbone for feature extraction. Following the convolutional layer of ResNet, an SPPF module is added to implement a feather-level fusion of local and global features. This process results in obtaining deep feature maps with resolutions of 1, 1/2, 1/4, 1/8, and 1/16 times of the original image. The channel numbers for these feature maps are 64, 128, 256, 512, and 512, respectively (referred to as $d1$, $d2$, $d3$, $d4$, and $d5$ in Fig. 1). Among them, the high-level deep features contain rich semantic information and have a large receptive field, while the low-level features have a small receptive field but contain detailed information. To fuse these features effectively, jump connections are employed to combine the high-level deep features with the low-level features.

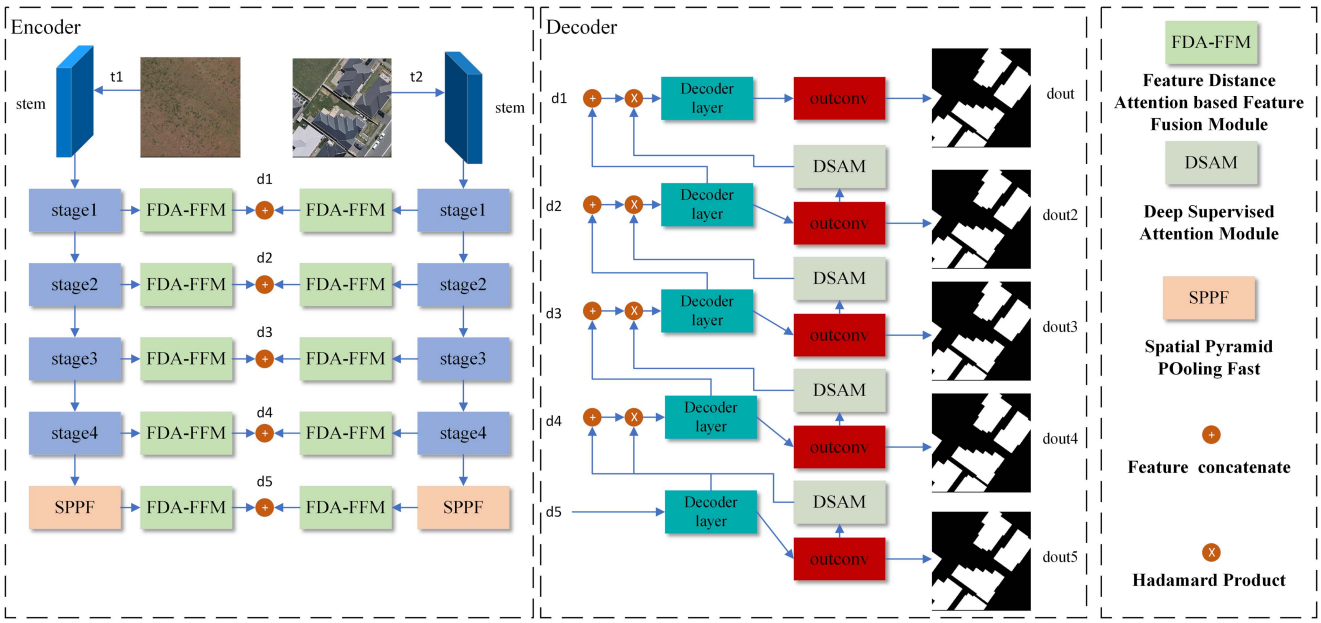


Fig. 1. FDA-FFNet framework.

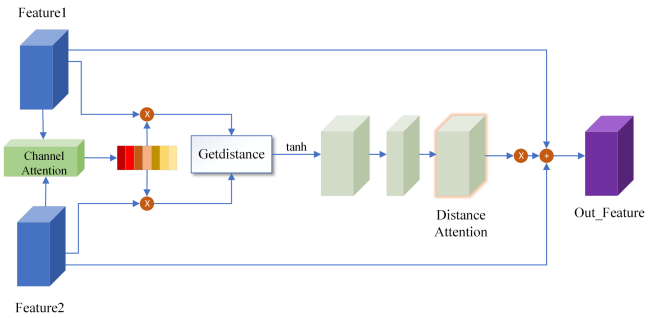


Fig. 2. Illustration of the proposed FDA-FFM.

B. Feature-Difference Attention-Based Feature Fusion Module

To enhance the model's capability to focus on the potential areas of change in the image, a novel module called FDA-FFM is proposed. The structure of FDA-FFM is illustrated in Fig. 2. The module draws inspiration from the concept of calculating feature differences in machine learning, in which vectors are commonly employed to represent each sample. The similarity of vectors can then be calculated to quantify the differences between the sample vectors. There are three primary methods for measuring the similarity of vectors: Euclidean distance, cosine distance, and Hamming distance. Among these methods, Euclidean distance is considered the most versatile. Normalized Euclidean distance is an improvement over the shortcomings of Euclidean distance. The presence of varying scales among the dimensions of the data can result in different Euclidean distance outcomes, introducing errors in determining the similarity between vectors. To address this, each dimension is individually processed using standardization, which involves adjusting the dimensions to have a mean of zero and a variance of one. This ensures that each dimension follows a standard normal distribution. The formula for calculating the standardized Euclidean distance is given as follows:

$$d = \sqrt{\sum_{k=1}^{n} \left(\frac{x_{1k} - x_{2k}}{S_k} \right)^2} \tag{1}$$

where S_k represents the standard deviation in the kth dimension, x_{1k} represents the eigenvalue of the kth dimension of the first set of vectors, and x_{2k} represents the eigenvalue of the kth dimension of the second set of vectors.

However, in deep learning, features extracted by the encoder constitute a set of feature vectors for each pixel point. To determine the attention of the feature differences, we adopt the concept of standardized Euclidean distance. Two-period channel attention [34] is employed to assign weights to the different channels of the features. After weighting the features of each channel, the Euclidean distance calculation is performed to obtain the attention of feature differences. Since the distance values fall within the range of [0,+∞), the sigmoid function commonly used in the attention mechanism is not suitable. Instead, we utilize the tanh function within the normalization function to obtain the differential attention values ranging from [0,1]

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \; x \in [0, +\infty), \; \text{Tanh} \in (0, 1) \tag{2}$$

$$A_d = \text{conv}\left(\tanh \sqrt{\sum_{k=1}^{n} A_{ck} (F_{1k} - F_{2k})^2 + \exp} \right) \tag{3}$$

where A_{ck} represents the channel attention in the kth dimension of the feature, F_{1k} represents the feature matrix in the kth dimension of the first period image, represents the feature matrix in the kth dimension of the second period image, and the term

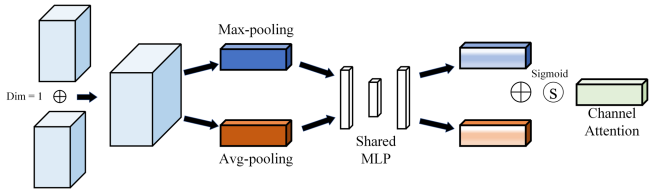


Fig. 3. Illustration of the proposed CCAM.

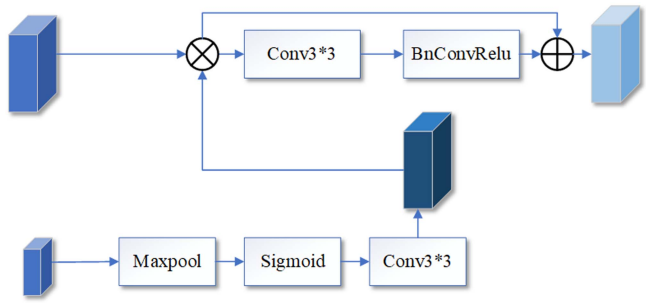


Fig. 4. Illustration of the proposed DSAM.

"exp" is a very small value, typically set to 0.00001, which is used to prevent division by 0 during gradient descent.

To ensure consistency in the feature weights of the same channel across bitemporal images, we make modifications to the channel attention mechanism (CAM) [34]. In our approach, we utilize the features extracted from both bitemporal images and perform max-pooling and avg-pooling operations. The subsequent modules of the CAM remain unchanged from the original implementation.

Considering that the distinguishing features of the bitemporal images have already been considered in the distance calculation, we employ a cascade operation for the fusion process. This involves calculating the difference attention and passing it through another spatial attention module [31]. The calculation process for obtaining the final fusion features is given as follows:

$$F_{\text{out}} = A_s\left(\text{Resblock} * 2\left(A_d\left(F_1 \hat{+} F_2\right)\right)\right) \tag{4}$$

where $\hat{+}$ represents a concatenation.

C. Deep Supervisory Attention Module

Feature pyramid networks [27] provide a well-established architecture, known as coarse-to-fine, for multilevel feature fusion. This architecture is commonly used in image segmentation network decoders. On top of that, a deep supervision strategy [32] has been proposed to generate and supervise results at each level. However, we believe that solely supervising the feature output at each level does not fully utilize the potential of features at each level. To address this, we have designed a deeply supervised attention module (DSAM) that better integrates the multilevel features and makes use of the deeply supervised features, as shown in Fig. 3.

D. Spatial Pyramid Pooling Fast

The SPPF module, illustrated in Fig. 5, proposed by Glenn Jochner [43] as an extension of SPP, offers significantly faster

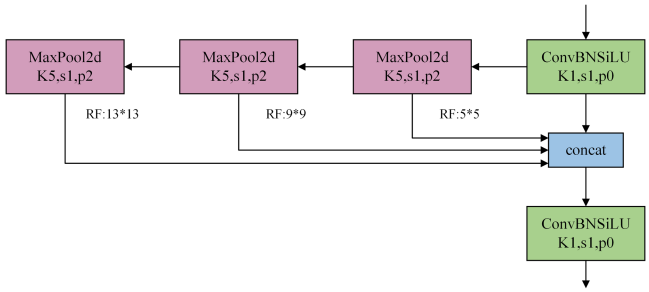


Fig. 5. Illustration of the proposed SPPF.

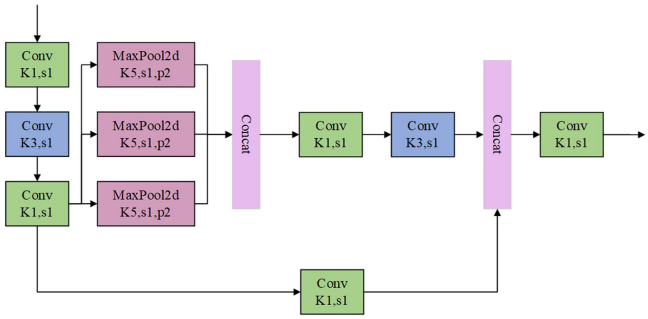


Fig. 6. Illustration of the proposed SPPCSPC.

processing speed compared with SPP [36] (2.5 times faster). In the YOLOV5 model, SPPF is utilized at the encoder end to implement the feature-map-level fusion of local and global features. However, in the latest YOLOv7 model [37], Wang et al. [36] proposed SPPCSPC, which outperforms SPPF in terms of performance. SPPCSPC draws on the ideas of CSPNet [38] and adds more residual connections to SPP, thereby improving the accuracy but also increasing the computation. The structure of SPPCSPC is illustrated in Fig. 6. Considering that the dataset for the CD task is typically smaller compared with object detection datasets, training an overly complex network might not yield optimal results. Hence, in our model, we utilize SPPF [43] after a ResNet encoder to extract the multiscale features of the image.

E. Loss Function

In the CD task, the scales of changed regions often vary, leading to a problem of scale imbalance. To alleviate this issue, we adopt a hybrid loss function that combines a binary cross-entropy (BCE) loss [39] and an IoU loss [40]

$$l = l_{\text{BCE}} + l_{\text{IoU}}. \tag{5}$$

The BCE loss function is a pixelwise loss function and is the most widely used loss in segmentation tasks. It weights all pixels equally and calculates the loss value for each pixel for the entire image as follows:

$$l_{\text{BCE}} = \sum_{x=1}^{H}\sum_{y=1}^{W} p_{xy} \ln g_{xy} + (1 - p_{xy}) \ln (1 - g_{xy}) \tag{6}$$

where g_{xy} and p_{xy} represent the true and predicted values, respectively.

TABLE I
QUANTITATIVE COMPARISONS IN TERMS OF PRE, REC, IOU, AND $F1$ ON TWO RSCD DATASETS

Dataset	FLOPs(G)	Params(M)	WHU-CD				LEVIR-CD			
			Pre(%)	Rec(%)	IoU(%)	F1(%)	Pre(%)	Rec(%)	IoU(%)	F1(%)
FC-Cat	10.62	1.55	72.26	84.52	63.81	77.91	87.80	89.42	79.54	88.60
FC-Diff	9.42	1.35	70.33	76.41	61.23	75.95	91.68	85.65	79.47	88.56
SNUNet	246.22	24.07	90.29	85.88	78.62	88.03	88.43	88.37	79.22	88.41
DSFIN	164.56	50.71	96.33	93.40	90.19	94.84	90.61	92.17	84.13	91.38
ChangeFormer	26.00	20.75	94.89	95.13	89.62	94.42	**92.97**	90.61	84.80	91.77
A2-Net	6.02	3.78	**97.23**	94.30	91.82	95.74	92.90	90.61	84.72	91.73
Ours_MobileNetv2	3.32	6.83	96.21	95.31	91.86	95.50	92.01	91.56	84.82	92.78
ours	209.54	53.79	96.70	**95.50**	**92.49**	**96.10**	92.13	**92.31**	**85.56**	**92.24**

The best results are highlight in bold entities.

The IoU loss function , a map-level measure, is integrated to focus on the overall detection accuracy of the change information and the global structural information. Its calculation formula is given as follows:

$$l_{\text{IoU}} = 1 - \frac{\sum_{x=1}^{H}\sum_{y=1}^{W}(g_{xy} * p_{xy}) + 1}{\sum_{x=1}^{H}\sum_{y=1}^{W}(g_{xy} + p_{xy} - g_{xy} * p_{xy}) + 1} \quad (7)$$

where g_{xy} and p_{xy} represent the true and predicted values, respectively.

When combining these losses, we utilize BCE loss to maintain a smooth gradient for all pixels while using IoU loss to give more focus on the foreground.

IV. EXPERIMENTS

A. Dataset

To evaluate the proposed method, we use two benchmark datasets for RSCD. The detailed information for each dataset is provided as follows.

1) *WHU-CD [41]:* A publicly available building CD dataset. It consists of a pair of optical very high resolution (VHR) RS aerial images with a size of 32 507 × 15 354 and a spatial resolution of 0.075 m. Similarly, we crop the original samples into small blocks of size 256 × 256 pixels. We randomly divided the dataset into a training set, validation set, and test set, consisting of 6096, 762, and 762 samples, respectively.
2) *LEVIR-CD [42]:* Consists of 637 pairs of Google Earth image patches with a VHR of 0.5 m/pixel. The size of each image patch is 1024 × 1024 pixels. These diachronic images span a period of 5–14 years and exhibit significant land use changes, especially related to the growth of the construction industry. LEVIR-CD covers a variety of building types, such as villa houses, high-rise apartments, small garages, and large warehouses. We divided each image pair into nonoverlapping patches of size 256 × 256 pixels, resulting in 3167 patches for training, 436 patches for validation sets, and 972 patches for testing.

B. Experimental Configuration

The proposed network is implemented using the PyTorch toolbox [44] and trained/inferred on a single Nvidia RTX 3090 GPU. We use Adam optimization [45] with a momentum of 0.9, weight decay set to 0.0001, and parameters $\beta 1$ and $\beta 2$ set to 0.9 and 0.999, respectively. The initial learning rate is set to 0.0005 and the batch size is set to 16. To enhance the robustness of the model, we leverage data augmentation techniques, such as random flipping, cropping, and temporal exchanging, to the input images. However, we disable data augmentation for the last 20 rounds of training.

C. Overall Comparison

We conducted a comparative analysis of the proposed model against six state-of-the-art RSCD methods, including seven CNN-based methods: FC-Diff [21], FC-Cat [21], SNUNet [29], DSIFN [27], ChangeFormer [31], and A2-Net [32].

Qualitative Evaluation: Table I presents a comparison of the quantitative evaluation result of the different methods on the two RSCD datasets, considering metrics, such as IoU, $F1$, Recall (Rec), and Precision (Pre). The results consistently demonstrate that the proposed method outperforms the existing methods. For example, on the LIVIR-CD dataset, our method achieves a 0.76% improvement IoU and a 0.41% improvement in $F1$. Similarly, on the WHU-CD dataset, our method achieves a 0.67% improvement in IoU and a 0.36% improvement in $F1$. Furthermore, the proposed method with light encoder MobileNetv2 also achieves excellent performance.

Qualitative Evaluation: The visual comparisons of different methods on the two RSCD datasets are shown in Fig. 6. We observe that the proposed method exhibits superiority in the following aspects.

1) Well-Defined Boundary: Compared with other methods, the proposed method can more accurately locate the boundaries of changed objects, and the recognition results have a more regular shape. In the first three rows of Fig. 6, FC-Cat, FC-Diff, SNUNet, DSFIN, ChangeFormer, and A2-Net struggle

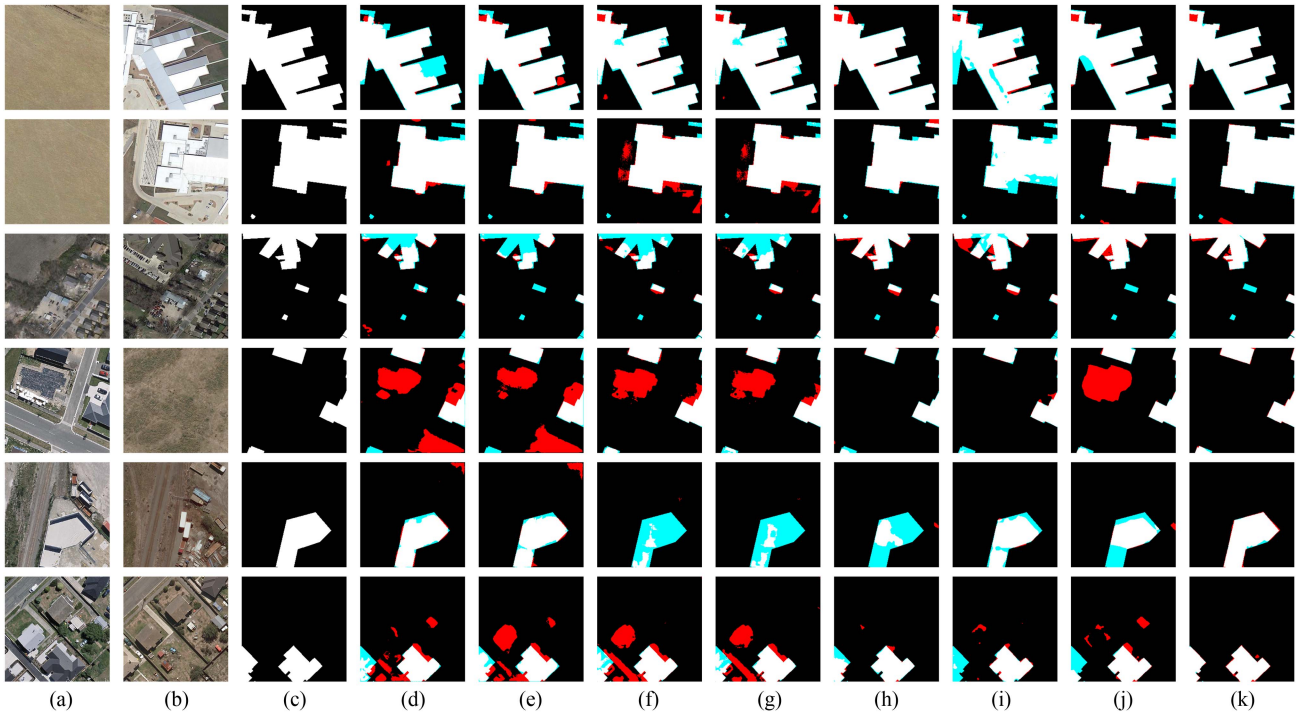


Fig. 7. Visual comparisons of the proposed method and the state-of-the-art approaches on the LEVIR-CD and WHU-CD dataset. (a) *t*1 images. (b) *t*2 images. (c) Ground truth. (d) FC-Siam-Conc. (e) FC-Siam-Diff. (f) SNUNet. (g) DSFIN. (h) ChangeFormer. (i) A2-Net. (j) Ours. The rendered colors represent true positives (white), false positives (red), true negatives (black), and false negatives (blue).

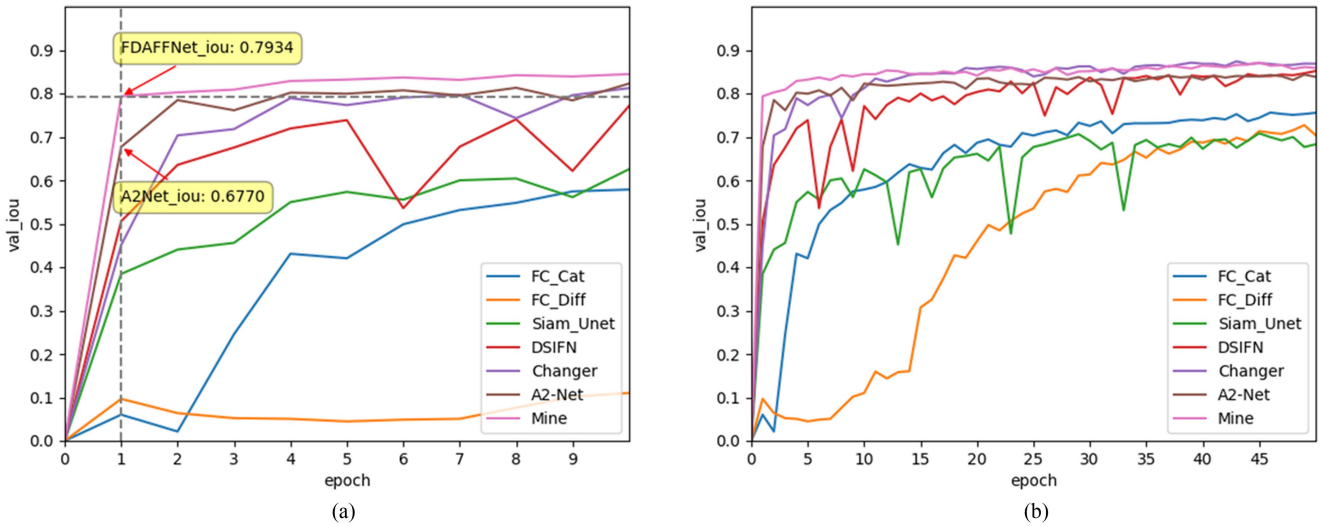


Fig. 8. Validation IoU of the proposed method and the state-of-the-art approaches on the LEVIR-CD in training process. (a) Magnified view of (b) during the first ten epochs. (b) Validation IoU in training process.

to accurately identify the boundaries of changing buildings, leading to numerous gaps in the detection results. In contrast, the proposed method effectively identifies potential change areas through FDA-FFM and further refines the extraction results at different levels using DSAM.

2) Better Distinguishing Pseudochanges: In the fourth and fifth rows of Fig. 7, the same buildings show different colors at different times, or nonbuilding areas undergo changes. This can potentially lead to misidentifications. Methods, such as FC-Cat, FC-Diff, STANet, SNUNet, A2-Net, and others, fail to correctly identify those erroneous changes. In the sixth row of Fig. 7, STANet, SNUNet, and DSFIN incorrectly classify the changed objects as nonbuildings, while other models also struggle to fully identify the changes. In contrast, the proposed method accurately distinguishes these pseudochanges.

3) Faster Convergence Speed: We monitor the training process of the proposed model and the proposed method and these state-of-the-art RSCD methods on the LIVIR-CD dataset. As shown in Fig. 8, the proposed method converges faster than other methods, achieving 79.34% IoU in just one epoch of training and an 11.64% improvement than the second method, A2-Net.

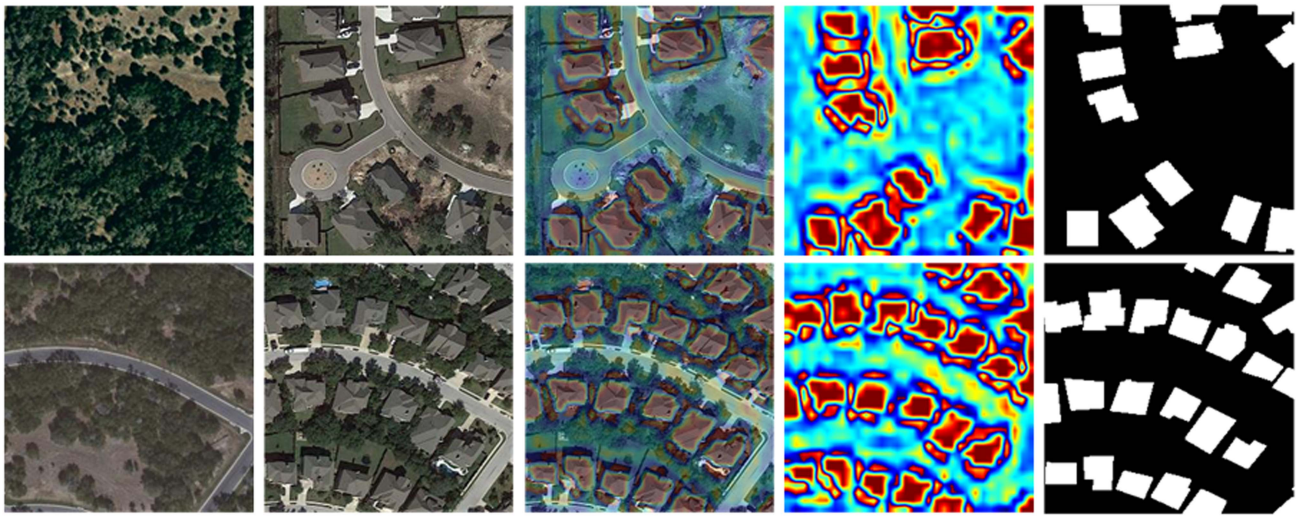


Fig. 9. Illustration of distance attention obtained by FDA-FFM. (a) $t1$ images. (b) $t2$ images. (c) Fusion image with $t2$ image and distance attention heatmap. (d) Distance attention heatmap. (e) Ground truth.

TABLE II
QUANTITATIVE COMPARISONS OF THE PROPOSED METHOD WITH DIVERSE SETTINGS IN TERMS OF PRE, REC, IOU, AND $F1$ ON THREE RSCD DATASETS

Dataset	WHU-CD				Levir-CD			
	Pre(%)	Rec(%)	IoU(%)	F1(%)	Pre(%)	Rec(%)	IoU(%)	F1(%)
Baseline	**96.70**	**95.50**	**92.49**	**96.10**	92.31	**92.06**	**85.56**	**92.19**
(a) FDA-FFM (FLOPs: 43.05G , Params 0.88M)								
Without_FDA	93.51	95.00	89.12	94.25	91.05	91.55	83.99	91.30
With_sig	96.49	94.92	91.76	95.70	91.39	91.57	84.29	91.47
(b) DSAM (FLOPs: 2.19G , Params 0.13M)								
Without_DSAM	96.20	94.75	91.33	95.47	92.26	91.48	84.96	91.87
(c) SPPM (SPPF: FLOPs:10.86G , Params 0.16M ; SPPCSPC: FLOPs: 116.55G , Params 1.77M)								
With_CSPC	94.70	94.89	90.11	94.74	91.79	91.63	84.69	91.71
Without_SPPF	95.01	95.06	90.54	95.03	92.06	91.75	85.02	91.90
(d) Backbone								
Resnet18	96.21	94.75	91.34	95.47	**92.61**	91.62	85.39	92.12
MobileNet	95.12	94.04	89.70	94.57	91.06	91.47	83.93	91.27
MobileNet_v2	96.21	95.31	91.86	95.50	92.01	91.56	84.82	92.78

*ALL FLOPs and Params are calculated in C, H,W = 256.
The best results are highlight in bold entities.

D. Ablation Studies

To validate the effectiveness of the components and configurations of the proposed network, we conducted a comprehensive ablation study on two RSCD datasets.

1) Effectiveness of FDA-FFM: The FDA-FFM aims to enhance the model's focus on potential change areas by calculating weighted Euclidean distances for features. To validate the effectiveness of FDA-FFM, we replaced it with a normal concatenation layer that has the same output channels as FDA-FFM. We formulated a method termed "Without_FDA" (i.e., (a) in Table II). Additionally, we evaluated the impact of using the sigmoid function instead of the tanh function in FDA-FFM. In the configuration termed "With_sig" (i.e., (a) in Table II), we replaced the tanh function with the sigmoid function. The quantitative comparison results are reported in Table II. As can be seen, OurNet outperforms OurNet-Without_FDA and OurNet-With_sig. OurNet-With_sig is only 0.30% over OurNet-Without_FDA.

The visual representation of the distance attention obtained by FDA-FFM is shown in Fig. 7. It can be observed that FDA-FFM effectively guides models to focus more on areas that likely undergo changes. From the heatmap in Fig. 9, the

changing buildings and the open spaces around them have been assigned high weights. Some roads have been assigned medium weights, while other areas have been given low weights.

2) Effectiveness of DSAM: DSAM aims to utilize the results of deep-supervised learning to refine CD results on a cascading basis. To validate the effectiveness of DSAM, we removed this module and formulated a method termed "Without_DSAM" (i.e., (b) in Table II). Removing DSAM leads to a degradation of approximately 1.16% and 0.64% in terms of IoU in the two datasets, indicating its significant contribution to detection accuracy.

3) Effectiveness of SPPF: SPPF aims to fuse local and global features at the feature-map level. To validate its effectiveness, we performed an ablation study by replacing SPPF with SPPCSPC and a normal convolutional layer, referred to as "With_CSPC" and "Without_SPPF." It can be seen that SPPCSPC actually decreases the model's accuracy by 0.33% in LEVIR-CD.

4) Others: We further validated the proposed method using different backbones, such as ResNet18 and MobileNet. From the results in Table II, we observed that the network with ResNet34 slightly outperforms the network with ResNet18 and performs better than the network with MobileNet on the LEVIR-CD and WHU-CD datasets. We argue that MobileNet is a lightweight network with poor feature extraction capability, and corresponding modules need to be designed to enhance its feature extraction capability.

V. Conclusion

In this article, we propose a feature-distance attention-based RSCD network for high-resolution RS image CD. In this method, we introduce FDA-FFM to enhance the model's capability to focus on potential areas of change in the image. Additionally, we incorporate SPPF to capture multiscale object information more efficiently, and we leverage DSAM to reconstruct change results from coarse to fine levels, utilizing deeply supervised outcomes. The experimental results demonstrate that the proposed method outperforms the current state-of-the-art CD methods. The method performs well in multiscale change information extraction accuracy, boundary extraction, and pseudochanges distinguishing. It has high reliability and practical application value. This work could serve as a new solution to feature fusion for RSCD.

References

[1] A. H. Chughtai, H. Abbasi, and I. R. Karas, "A review on change detection method and accuracy assessment for land use land cover," *Remote Sens. Appl., Soc. Environ.*, vol. 22, Apr. 2021, Art. no. 100482, doi: 10.1016/j.rsase.2021.100482.

[2] S. W. Wang, B. M. Gebru, M. Lamchin, R. B. Kayastha, and W.-K. Lee, "Land use and land cover change detection and prediction in the Kathmandu district of Nepal using remote sensing and GIS," *Sustainability*, vol. 12, no. 9, May 2020, Art. no. 3925, doi: 10.3390/su12093925.

[3] C. M. Viana, Sandra Oliveira, S. Oliveira, and J. Rocha, "29 - land use/land cover change detection and urban sprawl analysis," in *Spatial Modeling in GIS and R for Earth and Environmental Sciences*, H. R. Pourghasemi and C. Gokceoglu, Eds. Amsterdam, The Netherlands: Elsevier, 2019, ch. 29, pp. 621–651, doi: 10.1016/B978-0-12-815226-3.00029-6.

[4] S. W. Wang, L. Munkhnasan, and W.-K. Lee, "Land use and land cover change detection and prediction in Bhutan's high altitude city of Thimphu, using cellular automata and Markov chain," *Environ. Challenges*, vol. 2, Jan. 2021, Art. no. 100017, doi: 10.1016/j.envc.2020.100017.

[5] M. Huang and S. Jin, "Rapid flood mapping and evaluation with a supervised classifier and change detection in Shouguang using Sentinel-1 SAR and Sentinel-2 optical data," *Remote Sens.*, vol. 12, no. 13, Jan. 2020, Art. no. 2073, doi: 10.3390/rs12132073.

[6] S. Ma, H. Qiu, S. Hu, D. Yang, and Z. Liu, "Characteristics and geomorphology change detection analysis of the Jiangdingya landslide on July 12, 2018, China," *Landslides*, vol. 18, pp. 383–396, Jan. 2021, doi: 10.1007/s10346-020-01530-3.

[7] L. Wang, M. Zhang, X. Shen, and W. Shi, "Landslide mapping using multilevel-feature-enhancement change detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3599–3610, Feb. 2023, doi: 10.1109/JSTARS.2023.3245062.

[8] L. Wu et al., "Multi-type forest change detection using BFAST and monthly Landsat time series for monitoring spatiotemporal dynamics of forests in subtropical wetland," *Remote Sens.*, vol. 12, no. 2, Jan. 2020, Art. no. 341, doi: 10.3390/rs12020341.

[9] M. D. Negassa, D. T. Mallie, and D. O. Gemeda, "Forest cover change detection using geographic information systems and remote sensing techniques: A spatio-temporal study on Komto Protected forest priority area, East Wollega Zone, Ethiopia," *Environ. Syst. Res.*, vol. 9, no. 1, Jan. 2020, Art. no. 1, doi: 10.1186/s40068-020-0163-z.

[10] B. Yuan et al., "Spatiotemporal change detection of ecological quality and the associated affecting factors in Dongting Lake Basin, based on RSEI," *J. Cleaner Prod.*, vol. 302, Apr. 2021, Art. no. 126995, doi: 10.1016/j.jclepro.2021.126995.

[11] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000, doi: 10.1109/36.843009.

[12] P. Coppin and M. E. Bauer, "Digital change detection in forest ecosystems with remote sensing imagery," *Remote Sens. Rev.*, vol. 13, no. 3/4, pp. 207–234, Apr. 1996, doi: 10.1080/02757259609532305.

[13] R. D. Johnson and E. S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int. J. Remote Sens.*, vol. 19, no. 3, pp. 411–426, 1998.

[14] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS J. Photogramm. Remote Sens.*, vol. 61, no. 2, pp. 125–133, Nov. 2006, doi: 10.1016/j.isprsjprs.2006.09.004.

[15] S. Jin and S. A. Sader, "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances," *Remote Sens. Environ.*, vol. 94, no. 3, pp. 364–372, Feb. 2005, doi: 10.1016/j.rse.2004.10.012.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations, Comput. Biol. Learn. Soc.*, 2015, pp. 1–14.

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

[18] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv.1804.02767*.

[19] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv.1704.04861*.

[20] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv.2004.10934*.

[21] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067, doi: 10.1109/ICIP.2018.8451652.

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

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

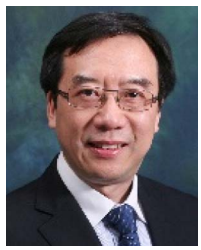
[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2018, pp. 3–11.

[25] X. Zhang et al., "DifUnet++: A satellite images change detection network based on UNet++ and differential pyramid," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 8006605.

[26] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020, doi: 10.1016/j.isprsjprs.2020.06.003.
[27] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5604816, doi: 10.1109/TGRS.2021.3085870.
[28] Y. Zhang, L. Fu, Y. Li, and Y. Zhang, "HDFNet: Hierarchical dynamic fusion network for change detection in optical aerial images," *Remote Sens.*, vol. 13, no. 8, Apr. 2021, Art. no. 1440, doi: 10.3390/rs13081440.
[29] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 8007805, doi: 10.1109/LGRS.2021.3056416.
[30] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 4401015, doi: 10.1109/TGRS.2023.3241257.
[31] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5610111, doi: 10.1109/TGRS.2023.3277496.
[32] Z. Li et al., "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5602812, doi: 10.1109/TGRS.2023.3241436.
[33] Q. Zhu et al., "Land-use/land-cover change detection based on a siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sensing*, vol. 184, pp. 63–78, 2022, doi: 10.1016/j.isprsjprs.2021.12.005.
[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[35] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, 2015, pp. 562–570.
[36] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.
[38] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 390–391.
[39] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, 2005, doi: 10.1007/s10479-005-5724-z.
[40] G. Máttyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3438–3446.
[41] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019, doi: 10.1109/TGRS.2018.2858817.
[42] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, May 2020, Art. no. 1662, doi: 10.3390/rs12101662.
[43] G. Jochner, "v7.0—YOLOv5 SOTA realtime instance segmentation," Code Ocean, 2022. [Online]. Available: https://github.com/ultralytics/yolov5/releases
[44] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 1–12.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

Wenguang Peng received the B.S. degree in photogrammetry and remote sensing in 2021 from Wuhan University, Wuhan, China, where he is currently working toward the M.S. degree in geographic information system.

His research interests include change detection, remote sensing scene classification, and deep learning for remote sensing.



Wenzhong Shi received the Ph.D. degree in geomatics from the University of Osnabrück, Vechta, Germany, in 1994.

He is currently the Director of PolyU-Shenzhen Technology and Innovation Research Institute (Futian), Shenzhen, China, the Director of Otto Poon Charitable Foundation Smart Cities Research Institute, PolyU, Hong Kong, Chair Professor of GISci and remote sensing, and the Director of Joint Research Laboratory on Spatial Information, PolyU, and Wuhan University, Wuhan, China. He is the Academician of the International Eurasian Academy of Sciences and a Fellow of the Academy of Social Sciences, London, U.K. He has authored or coauthored more than 300 research articles in journals indexed by SCI and 20 books. His research covers urban informatics for smart cities, GISci and remote sensing, specifically, AI-based object recognition and change detection from satellite imagery, intelligent analytics and quality control for spatial data, mobile mapping and 3-D modelling based on LiDAR and imagery, and 3-D GIS models.

Min Zhang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2020.

He is currently a Research Assistant Professor with the Department of Land Surveying and GeoInformatics, The Hong Kong Polytechnic University, Hong Kong. His research interests include spatial data quality, deep learning, artificial intelligence, change detection, and object recognition in remote sensing.

Lukang Wang received the B.S. degree in surveying engineering in 2018 from China University of Mining and Technology, Xuzhou, China, where he is currently working toward the Ph.D. degree in geodesy and surveying engineering with the School of Environment and Spatial Informatics.

His research interests include change detection, landslide mapping, and deep learning for remote sensing.