SASiamNet: Self-Adaptive Siamese Network for Change Detection of Remote Sensing Image

Xianxuan Long, Wei Zhuang, Min Xia, *Member, IEEE*, Kai Hu, and Haifeng Lin

***Abstract*—With increasingly rapid development of convolutional neural networks, the field of remote sensing has experienced a significant revitalization. However, understanding and detecting surface changes, which necessitate the identification of high-resolution remote sensing images, remain substantial challenges in achieving precise change detection. Excited deep learning-based change detection techniques often exhibit limitations and lack the necessary precision to detect edge details or other nuanced information in remote sensing images. To address these limitations, we propose a unique semantic segmentation deep learning network, the self-adaptive Siamese network (SASiamNet), specifically devised for enhancing change detection in remote sensing images. The SASiamNet excels in real-time land cover segmentation, adeptly extracting local and global information from images via the backbone residual network. Furthermore, it incorporates a primary feature fusion module to extract and fuse the primary stage feature map, and a high-level information refinement module to refine the resultant feature map. This methodology effectively transmutes low-level semantic information into high-level semantic information, thereby improving the overall detection process. Aimed at empirically testing the effectiveness of the SASiamNet, we utilize two distinct datasets: the public dataset, LEVIR-CD, and a challenging dataset, CDD. The latter is composed of bitemporal images sourced from Google Earth, spanning various regions across China. The experiment results unequivocally demonstrate that our approach outperforms traditional methodologies as well as contemporary state-of-the-art change detection techniques, hence underscoring the efficacy of the SASiamNet in the context of remote sensing image change detection.**

***Index Terms*—Change detection, deep learning, remote sensing, Siamese network.**

I. Introduction

CHANGE detection in remote sensing involves identifying semantic and nonsemantic changes in images. To be more specific, the related images are captured at different times from the same region. Recent hardware advancements have made acquiring these images easier, but precision (PR) in high-resolution image change detection remains challenging because of factors, such as lighting changes, atmospheric disturbances, and registration accuracy. Conventional and deep learning algorithms are proposed for change detection, but shortcomings persist.

Recent progress in hardware technology, particularly in sensor capabilities, has facilitated the acquisition of remote sensing imagery. Revisiting data provide more comprehensive land cover information compared with dense time series images. In addition, these images are with medium spatial resolution. However, this technological advancement presents its own challenges [1]. High-resolution images often suffer from a decline in registration accuracy, making precise change detection an intricate task [2]. In addition, dynamic variables, such as alterations in illumination and atmospheric disturbances, compound the complexity of obtaining sufficient feature information from bitemporal remote sensing images [3]. There are numerous of existing change detection methodologies, which can be mainly divided into two types: traditional techniques with conventional strategy to deal with change detection tasks and deep learning-based techniques. Traditional techniques, while valuable, demonstrate limitations when dealing with high-resolution data and complex environmental conditions [4], [5]. Deep learning-based techniques, although innovative and promising, still grapple with challenges in detecting detailed nuances, such as edge information. This article proposes a state-of-the-art approach that strives to address these constraints, thereby improving the PR in change detection of remote sensing image.

Generally, traditional techniques incorporate the image arithmetic method, whose main strategy is comparing the pixel values of two bitemporal images, and then producing a differential image. By establishing a suitable threshold, pixels are classified to indicate whether a region in an image is changed or unchanged. Among traditional techniques for image change detection, principal component analysis (PCA) is commonly used, which is a classical algorithm for dimensionality reduction based on image transformation [6]. However, the dependency of PCA on the statistical properties of data exposes it to considerable impacts arising from unbalanced datasets. In response to this problem, Celik [7] introduced an unsupervised change detection technique, which effectively combines PCA with K-means clustering. Singh and Singh [8] proposed a technique, which considers both spectral and statistical properties, utilizing canonical variate analysis (CVA) to seize spectral change information and a fuzzy classifier for change detection in fused images.

Manuscript received 27 July 2023; revised 4 October 2023 and 23 October 2023; accepted 28 October 2023. Date of publication 6 November 2023; date of current version 6 December 2023. This work was supported in part by the National Natural Science Foundation of PR China under Grant 42075130 and in part by the Postgraduate Research and Innovation Project of Jiangsu Province under Grant 1534052101072. *(Corresponding author: Wei Zhuang.)*

Xianxuan Long, Wei Zhuang, Min Xia, and Kai Hu are with the Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, B-DAT, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: 202083710036@nuist.edu.cn; zw@nuist.edu.cn; xiamin@nuist.edu.cn; 001600@nuist.edu.cn).

Haifeng Lin is with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210000, China (e-mail: haifeng.lin@njfu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3330753

Despite the fact that unsupervised methods anchored on arithmetic and transformation lack prior knowledge of labeled data, and primarily hinge on model assumptions or comparative rules to pinpoint changed areas, they are afflicted by certain limitations. To enhance change detection performance, researchers have shifted from unsupervised to supervised approaches. Deng et al. [9] proposed a method that can effectively identify and quantify changes in land use through using hybrid classifiers to combine unsupervised and supervised classification. However, arithmetic and transformation-based methods heavily depend on empirical design, which makes them less effective for high-resolution images. Juan et al. [10] proposed a detection technique, integrating both pixel and object representations, which utilizes supervised subimages to discern areas of change that encompass artificial objects, achieved by partitioning large-scale images into overlapping subimages. Despite the advancements these methods have contributed to change detection, the challenge of accurately capturing texture characteristics in relatively complicated topographic environment and details in high-resolution images persists.

Deep learning algorithms have shown great potential in change detection in remote sensing images [11], [12], [13]. The inherent advantage of deep learning methods is their adeptness at recognizing complex features in images full of subtle semantics through hierarchical structures [14], [15], [16]. Compared with traditional change detection algorithms, deep learning methods do not require tedious manual design and show significant advantages in adapting to the various intricate features of complex remote sensing datasets. To be more specific, deep learning methods address issues, such as edge blurring and missed small targets in change detection, by leveraging their ability to learn complex hierarchical features directly from data. Recently, deep convolutional neural networks (CNNs) have played a significant role in the segmentation of remote sensing images and have continuously shown their efficacy in situations where change detection is required [17], [18], [19]. CNNs, renowned for their capacity to discern spatial hierarchies, learn from local input patches, enabling the precise demarcation of regions of change while preserving spatial continuity. Furthermore, the robust handling of multispectral data by CNNs, facilitated by their multilayered convolution process, allows for the extraction of features that effectively capture spectral, spatial, and temporal dynamics inherent to remote sensing imagery. These abilities are critical for the comprehensive identification and interpretation of changes over time. There exist many networks known for their outstanding performance in image classification tasks, such as the visual geometry group (VGG) network [20], residual network (ResNet) [21], deep learning with depthwise separable convolutions (Xception) [22], and the densely connected convolutional network (DenseNet) [23]. In the domain of remote sensing change detection, the inability of shallow networks to comprehensively extract image feature information from high-resolution images has led to a diminished standing. Ideally, to extract richer image information, a network should be as deep as possible. However, an increase in network layers could introduce the issue of the vanishing gradient, leading to the loss function nearing zero, subsequently resulting in decreased network training efficiency. After a thorough evaluation, we have selected ResNet as the backbone network for our proposed model due to its superior ability in mitigating the problem of gradient vanishing through its multitude of residual blocks. The residual block excels in fitting the classification function, thus achieving higher classification accuracy. It can effectively address the problem of optimization training as the network layers deepen. However, fitting a potential identity mapping function, $H(x) = x$, can be challenging for certain layers, potentially contributing to the complexity of training deep networks. Designing a network as $H(x) = F(x) + x$ is a simpler approach. Transitioning to learn a residual function, $F(x) = H(x) - x$, makes fitting the residual more straightforward once $F(x) = 0$ is achieved, forming an identity map, $F(x) = H(x) - x$. This allows the ResNet to eliminate identical components and emphasize the minor changes in the feature image. The primary aim is to extract sufficient feature information and combine low-level and high-level semantic features using an attention mechanism to improve prediction accuracy. General CNNs have been widely used in various computer vision tasks, including image classification and segmentation. They excel at learning hierarchical features from data, enabling the identification of complex patterns and structures in images. However, in the domain of remote sensing change detection, particularly with high-resolution imagery, the limitations of shallow networks become evident. Shallow networks struggle to comprehensively extract feature information from these images, which often exhibit subtle changes and intricate details. As a result, deep learning methods have emerged as a more effective solution, as they can address issues, such as edge blurring and the detection of small targets, while maintaining spatial continuity, making them particularly valuable in this context. As for recent methods proposed for remote sensing image change detection, Daudt et al. [24] introduced a trio of fully CNN architectures, of which two are extensions of the fully convolutional paradigm with Siamese configurations. This Siamese architecture has manifested superior performance in the domain of change detection when compared with antecedent methodologies. Chen and Shi [25] proffered a spatiotemporal attention neural network predicated upon connective principles, while Chen et al. [26] proposed dual-time image Transformers, adept at effectively modeling both temporal and spatial contexts. In a similar vein, Ding et al. [27] conceived a novel dual-branch end-to-end network tailored to the exigencies of building change detection. Notably innovative in its approach, this network introduces cross-layer addition and skip connection modules, judiciously guided by a spatial attention mechanism, thereby enabling the aggregation of multilevel contextual information. This innovation is instrumental in enhancing both network performance and robustness through the incorporation of deep supervision modules. Further advancing the domain, Wang et al. [28] devised a high-resolution feature difference attention network dedicated to the task of change detection. This network introduces a multiresolution parallel architecture, which comprehensively harnesses image information across varying resolutions to mitigate the degradation of spatial information. Notably, the intrinsic challenges within this realm are exacerbated by variations in sun altitude angles, lighting conditions, seasonal fluctuations, and the occlusive effects of building shadows across the two images. Consequently, identical

objects may exhibit stark disparities in their spatial positioning and spectral characteristics, giving rise to deviations both in position and spectral attributes. The complications are further magnified in the context of high-resolution images, rendering the task of change detection particularly arduous. Presently, extant deep learning-based change detection algorithms are faced with the formidable challenge of accurately discerning between areas that have undergone change and those that remain unaltered.

To solve the problems mentioned, our research propose self-adaptive Siamese network (SASiamNet) for high-resolution remote sensing image change detection. In our study, we create two auxiliary modules to assist network training. To be more specific, to distinguish and classify the changed area and unchanged area, we introduce the difference module (DM) and assimilation module (AM), respectively, which are proven to have an outstanding performance in extracting features of bitemporal remote sensing images. Moreover, with another module called representative feature extraction module (RFEM), we focus on extracting the temporal features of two remote sensing images. Considering combining the features extracted by these three submodules, a primary feature fusion module (PFFM) is created to fuse the output of DM, AM, and RFEM. We also propose a module called high-level information refinement module (HIRM), which is aimed at aggregating and refining the global information from different scales, especially for the high-level feature, so that we can attain more precise semantic information. These contributions form our main achievements in this study.

1) We introduce a state-of-the-art deep learning network, the SASiamNet, for high-resolution remote sensing change detection. Our network efficiently extracts pertinent image features, addressing issues, such as ambiguous changed target edges and small target omission. As an end-to-end trainable network, SASiamNet simplifies the change detection task by obviating the need for separate component training.
2) We introduce several modules for enhanced change detection, including the DM, RFEM, AM, PFFM, HIRM, and global feature fusion module (GFFM). DM and AM perform weight training on changed areas, augmenting RFEM to improve prediction accuracy across multiple feature map scales. HIRM, inclusive of the GFFM submodule, uses four feature map scales to refine and fuse predictions through adaptive weight allocation and effective information integration from varying feature map sizes.
3) We evaluate the proposed SASiamNet on two datasets: the publicly available LEVIR-CD dataset and our proposed CDD dataset. The results demonstrate the superior reliability and accuracy of SASiamNet compared with various existing deep learning methods for change detection.

II. Methodology

This study utilizes CNNs due to their process in semantic segmentation, particularly suited to classifying "changed" and "unchanged" regions in remote sensing images. To further enhance this approach, a Siamese network architecture is employed, taking advantage of its weight-sharing mechanism. Siamese networks, composed of two or more identical subnetworks, ensure a consistent parameter set for feature extraction from different images. This uniformity promotes a robust comparison between image pairs, a pivotal aspect of the change detection task. This weight-sharing characteristic allows the network to learn a similarity measure between paired inputs, leading to the extraction of more robust feature representations by observing the same changes from different perspectives or timepoints. The combination of CNNs' superior semantic segmentation capabilities and the Siamese networks' weight-sharing attribute provides a compelling approach for accurate and consistent change detection in remote sensing images.

The U-Net architecture serves as the foundation for our proposed network due to its structure, originally designed for semantic segmentation tasks. Its fully convolutional nature, combined with an encoder–decoder pattern, is advantageous for change detection in remote sensing images. The encoder captures the context while reducing spatial dimensionality, and the decoder uses these features to restore spatial dimensions, producing high-resolution output. Skip connections in U-Net enable the transfer of fine-grained details from encoder to decoder, thus aiding the generation of precise segmentation maps, which is pivotal for intricate change detection. The U-Net architecture's efficiency in training and its ability to perform well even with fewer training images enhances its suitability for our task. Furthermore, we selected ResNet [21] as the backbone of our architecture due to its demonstrated success in deep learning tasks. ResNet's introduction of residual learning addresses the vanishing gradient problem often encountered with deep networks. Through skip or shortcut connections, gradients can backpropagate to earlier layers without significant degradation, facilitating the learning of complex hierarchical features essential for effective change detection in remote sensing tasks. The wide acceptance and excellent performance of ResNet across various image recognition tasks further validate our selection for the network backbone. Moreover, two auxiliary modules are created: PFFM and HIRM. PFFM can effectively extract the feature of images in the primary stage, which means that it detects the difference and similarity between two remote sensing images by automatically positioning the pixels, and more importantly, it depresses the features that are relatively insignificant and less typical, and combines the output together to obtain a rather comprehensive feature map with rich information, enabling the whole network to effectively detect the small objects. More specifically, the submodule within PFFM, RFEM contributes a part to handle the issue of edge blurring and missing targets through enhancing the weight of representative features. The HIRM can aggregate the features from multiple scales and refines the semantic features effectively, which enables the network to fuse the high-level information and obtain more accurate semantic change information. More importantly, this module can adaptively select the features with more significant and representative semantic information, which is derived from the pixels positioned in PFFM. At this stage, through combining the

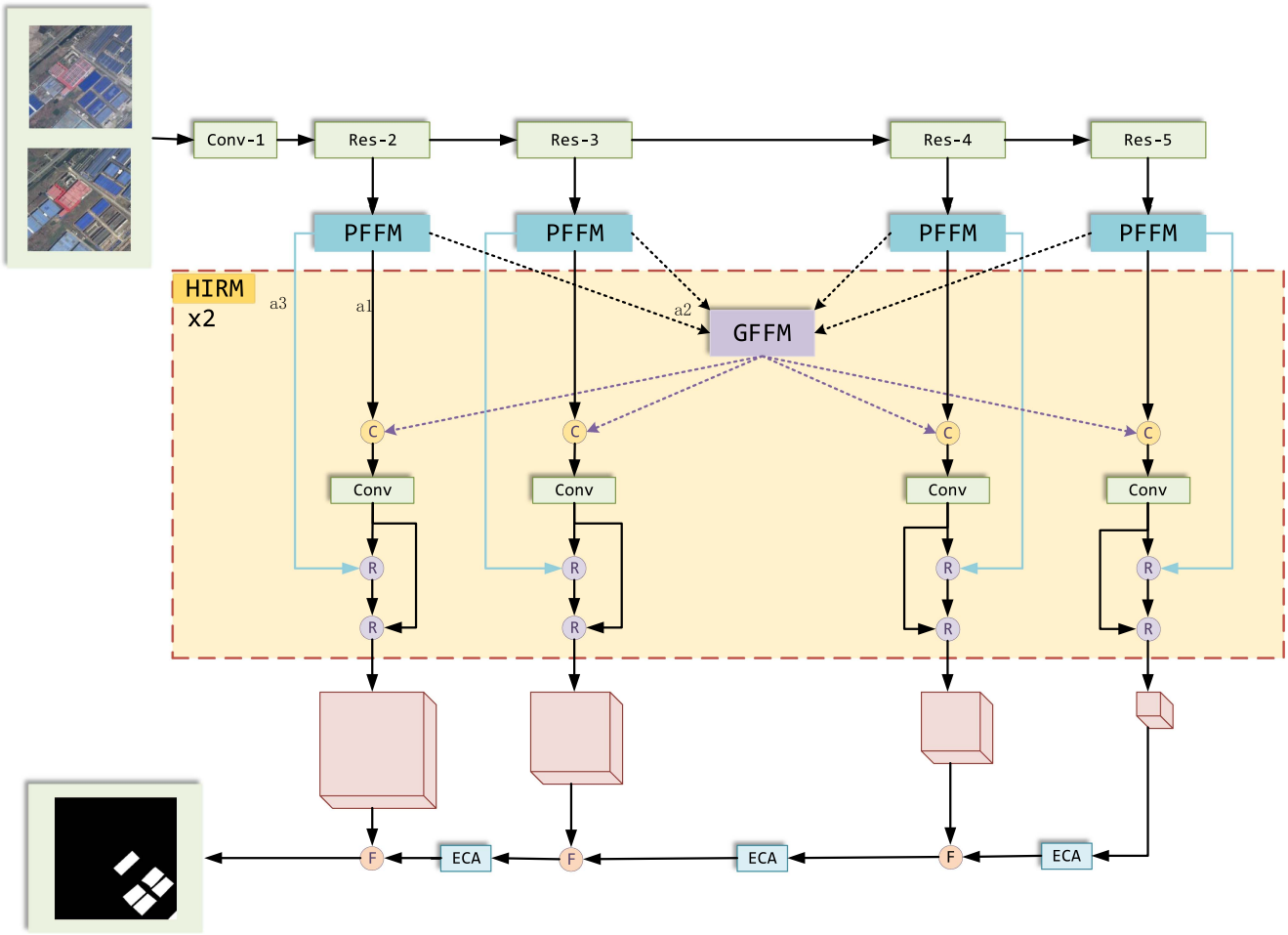

Fig. 1. SASiamNet framework: C represents concatenation, R represents feature map refinement, F represents fusion. Two auxiliary modules: PFFM, HIRM.

two auxiliary modules in our proposed method, the problem of misdetection is solved. The overall model structure is displayed in Fig. 1.

Given our intention to extract both deep and shallow information via the backbone, we have chosen to employ ResNet as our backbone network. Within this network, the backbone consists of five convolutional blocks: Conv-1, Res-2, Res-3, Res-4, and Res-5. We allow Conv-1 to remain detached, while integrating the remaining blocks, specifically Res-2 to Res-5, into the network, with the outputs from each layer serving as inputs for subsequent modules.

A. *Primary Feature Fusion Module*

The fundamental objective of an image change detection algorithm is to discern the areas of alteration via bitemporal images. The task of change detection primarily incorporates two interconnected subtasks: identifying regions of change and those of nonchange. Currently, a plethora of strategies exist to approach this task. Traditional remote sensing change detection algorithms, such as the improved ratio median absolute deviation (IR-MAD) method, MAD method, change vector analysis (CVA), and PCA are frequently used for change detection in remote sensing data. However, these techniques may prove ineffective in detecting minute changes or may exhibit sensitivity to data noise. With the rapid advancement of graphics processing units (GPUs) capable of handling an extensive volume of data and executing complex tasks with superior PR and speed compared with traditional central processing units, the efficiency of deep learning methods in complex scenery change detection has seen significant improvements. These methods include CNNs, recurrent neural networks, and generative adversarial networks. Nevertheless, current modules struggle to accommodate variability in lighting conditions, changes in weather, and differing shooting angles, which significantly contribute to errors in remote sensing detection. In response, we developed a robust module dubbed the PFFM designed to extract primary stage features and effectively fuse them. These primary stage features predominantly encapsulate low-level characteristics, which depict basic and local patterns or structures in the data, such as edges, colors, and textures. Importantly, this module exhibits resilience against a variety of environments and mitigates the influence of factors that could cause minor deviations in the results of remote sensing change detection. In essence, PFFM comprises three branches, each with three child modules designed to address distinct issues. The module's

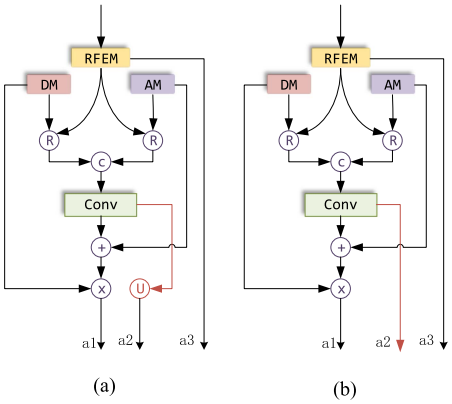

Fig. 2. (a) PFFM structure of the first branch, which is connected with the Res-2 layer. (b) PFFM structure of the rest four branches. R represents refinement, C represents concatenation, + represents elementwise summation, × represents elementwise multiplication, and U represents an upsampling operation.

structure is illustrated in Fig. 2. It is worth noting that the structure of the first branch, where the input originates from the Res-2 layer, slightly deviates from the other four branches due to the size of the feature map.

The inputs of each layer are different. As we design, considering the original image size is $C \times H \times W$, then the output of Res-2, Res-3, Res-4, and Res-5 are expected to be $64 \times \frac{1}{4}H \times \frac{1}{4}W$, $128 \times \frac{1}{8}H \times \frac{1}{8}W$, $256 \times \frac{1}{16}H \times \frac{1}{16}W$, and $512 \times \frac{1}{32}H \times \frac{1}{32}W$ respectively. More importantly, to avoid losing detail of feature maps, we generate three outputs from different location in this module, which could also be helpful for the entire network to get global semantic information. One output is the output of RFEM, which is denoted as $a3$. In this module, we design three branches: DM, AM, and RFEM. In order to extract the rich semantic information, we design to combine and refine the output feature map of DM and RFEM, and AM and RFEM, through adding them together, then passing the added feature map through a 2-D convolution layer, applying batch normalization and ReLU to the output, which also converts three branches to two branches, reducing the number of parameters. After feature refinement, the information of changed and unchanged regions is combined. Then, we concatenate the refined information and then send the output to a 2-D convolution layer. Similarly, we still apply a batch normalization and a ReLU. The reason why we always use the strategy of Conv+BN+ReLU is that convolutional layer extract feature information and can filter the information, and the implementation of batch normalization is responsible for smoothing the loss function and gradient decrease, and utilizing ReLU prevents the exponential rise of the compute needed to run the neural network. We also add the output from the main path in this module with the output of AM, then multiply the output with the output of the DM. Similarly, after each operation, such as add and multiplication, a BN layer and a ReLU layer are followed. The output of the main path in PFFM is denoted as $a1$. Considering combining the feature maps of different scales, we apply upsampling to the output from the branch of Res-3, Res-4, and Res-5, to make sure the image sizes are the same as the output from Res-2, where the output of upsampling is denoted as $a2$. Due to the size of the output from Res-2 is already the desired size, we do not employ upsampling to it.

1) Difference Module and Assimilation Module: The core of image change detection is identifying the modified areas between two bitemporal images. This endeavor is chiefly delineated by the tasks of detecting areas of change and areas of continuity. Accordingly, we have constructed a differentiation module and an AM. Each of these is tasked with extracting feature maps from the changed and unchanged areas, respectively. The inherent complexity of this undertaking is attributable to the variability of objects under different sensor perspectives. This results in a lack of one-to-one correspondence between all pixels in the two images, captured at distinct points in time. This predicament is accentuated in the presence of multisensor usage, which can lead to varied building angles, particularly for buildings with significant three-dimensionality. As an illustration, transformation detection in intricate urban scenarios—encompassing shifts in pedestrian movement, vehicles, and vegetation precipitated by seasonal weather variations and other elements—becomes increasingly vulnerable to minute deviations in sensor perspectives. This discrepancy becomes glaringly apparent when the observation distance is minimal, the object height is substantially high, and the spatial resolution is expansive, thereby preventing a one-to-one pixel correspondence. Conventional techniques, such as PCA-means and IR-MAD1, are predominantly unsupervised, lacking prior knowledge of labeled data and instead depending heavily on particular assumptions or similarity rules to discern changes. Furthermore, the majority of these conventional change detection methods necessitate module adjustments in the face of evolving scenarios, a process that is both time and labor-intensive. Conversely, the employment of unsupervised learning in the analysis of bitemporal remote sensing images may render the process excessively simplistic. Consequently, it becomes a formidable task for standard image processing technology to contend with change detection problems in complex scenarios. However, deep learning methods have demonstrated their capacity to effectively address these challenges, facilitating the conversion of area parameters into individual values for comparison. Existing modern techniques, such as SegNet and ENet, encounter constraints when processing high-resolution input images. Specifically, their lightweight base models risk compromising spatial information, resulting in considerable loss of the original image's spatial attributes. In response to this limitation, we put forward two feature extraction strategies that not only preserve the wealth of spatial information but also enable the detection of differences and similarities between two images of the same subject captured at separate intervals.

In the DM, the feature maps of the identical area captured at different time intervals are processed through the recursive residual group (RRG). Following this, a subtraction operation is performed on these feature maps, generating a new feature map that delineates the discrepancies between the two images. Conversely, the AM adds together the two feature maps, creating a feature map that emphasizes the similarities between the two images. The structures of DM and AM are depicted in Fig. 3. Both these modules are capable of transforming regional

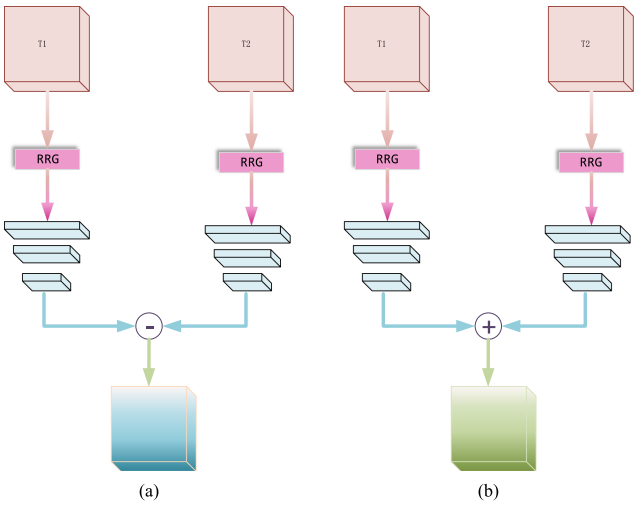

Fig. 3. (a) Structure of the DM. (b) Structure of the AM. RRG represents the recursive residual group, and - and + stands for elementwise subtraction and summation, respectively.

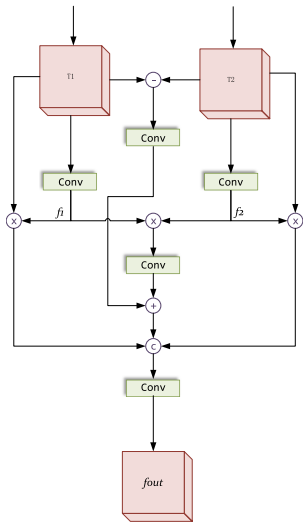

Fig. 4. Representative feature extraction module (RFEM).

parameters into single comparable values. To elaborate, in DM, once the feature map is processed by the RRG module—which maintains the size of the feature map—it is passed through three convolution layers, where the feature map's size reduces to one-eighth of the original dimension. Each convolution layer comprises a convolution operation with stride $= 2$, followed by batch normalization and a ReLU activation function. This method, which maps the information corresponding to the original feature image into a single value after several convolution blocks, significantly enhances the accuracy of our proposed network. These two modules serve unique roles within the proposed model. While the difference network emphasizes feature extraction from the altered region, the assimilation network concentrates on retrieving feature information from the unchanged region. The primary network is fused with the two auxiliary networks utilizing a multiscale feature aggregation method, forming the complete network. Throughout the training phase, the network parameters are updated by monitoring the training progress, enabling the model to learn all information autonomously. We establish a loss function to facilitate training. The functions of the two auxiliary networks can be better comprehended by visualizing their outputs on the original map.

2) Representative Feature Extraction Module (RFEM): CNNs are the most widely used deep learning models for extracting rich feature information. CNNs have proven to be extremely effective in many computer vision tasks, such as object recognition, image segmentation, and scene understanding. The ability of CNNs to automatically learn features from raw input data by applying filters at multiple scales allows them to capture hierarchical and complex features. This enables them to extract and represent relevant and discriminative features for various applications. In previous research, although some models (such as VGG, ResNet, InceptionNet, AlexNet, and DenseNet) can extract rich feature information from a feature map, the issue of feature redundancy emerges as the number of convolutional blocks increases. To be more specific, with the deepening of the CNN, the extracted feature map may contain some irrelevant features, which can also be seen as feature redundancy. Sometimes, the representative features are mixed with irrelevant features, leading to worse network performance, which increases the chance of the problem of misdetection. The representative feature stands for the features that CNN should learn to identify including changes in land cover, such as the conversion of green spaces to built-up areas, variations in infrastructure, such as the construction of new roads or buildings, and alteration in color pattern, such as vegetation, rooftops, or paved surfaces. As for irrelevant features, the dataset may also contain irrelevant features that do not contribute to accurate change detection, such as cloud cover, seasonal changes in vegetation, shadows, or differences in image acquisition conditions (e.g., lighting and sensor angle). On account of mixed features, misdetection may occur. More specifically, if CNN learns to associate some of these irrelevant features with change detection, it may lead to misdetection. For instance, the network may mistakenly learn that cloud cover in the T_2 image is an indicator of urban expansion, as it might obscure the actual land use changes. As a result, when the network encounters a pair of images where the only difference is the presence of clouds in the T_2 image, it may incorrectly identify urban expansion. CNN's performance is negatively impacted by the mixture of representative and irrelevant features during training. Having considered that some features are redundant, we propose a RFEM. This module mainly extracts the detailed feature of images and has the capability of enhancing the features that are crucial and discarding some unimportant features, which is shown in Fig. 4.

In RFEM, we first send the pair of the original feature maps to a 3×3 convolution layer, obtaining a pair of processed feature maps, denoted as f_1 and f_2, respectively, which are multiplied together and sent to another 3×3 convolution layer where the result is denoted as f_m. In this process, we denote the original input feature maps as f_{T1} and f_{T2}, then we can get the processed feature maps f_1 and f_2

$$f_1 = f^{3\times 3}(f_{T1}) \tag{1}$$

$$f_2 = f^{3\times3}(f_{T2}) \tag{2}$$

$$f_m = f^{3\times3}(f_{T1} \otimes f_{T2}). \tag{3}$$

In these formula, $f^{3\times3}(\cdot)$ represents the 2-D convolution, batch normalization, and ReLU activation function with convolution kernel size equaling 3, stride equaling 1, and padding equaling 1.

Then, we subtract the two original input feature maps of the same remote sensing images captured in different times and pass through a 3×3 convolution layer, which can be denoted as f_s

$$f_s = f^{3\times3}(\text{abs}(f_{T1} - f_{T2})) \tag{4}$$

where $\text{abs}(\cdot)$ denotes absolute difference operation.

Next, on account of the intention of reaching a balance between the rich feature information and the redundant information, we add the result of the subtracted feature map f_s and the multiplied feature map $f_s m$ together, obtaining a feature map f_a with strengthened difference features, which can be seen as the major branch of this proposed module

$$f_a = f_s \otimes f_m. \tag{5}$$

Then, we multiply the pair of the original feature maps with the processed feature maps f_1 and f_2, which are the two branches of RFEM. The result feature maps can be denoted as B_1 and B_2

$$f_{B1}w = f_{T1} \otimes f_1 \tag{6}$$

$$f_{B2} = f_{T1} \otimes f_2. \tag{7}$$

Finally, given that we intend to focus on the more representative information and also the global information, we concatenate the outputs of three branches and pass through a 1×1 convolution layer to adjust the channel number of the output. Thus, the output of RFEM can be represented as f_{out} in the following formula:

$$f_{\text{out}} = f^{1\times1}(\text{Concat}(f_a, f_{B1}, f_{B2})) \tag{8}$$

where $f^{1\times1}(\cdot)$ denotes a block with the 2-D convolution, batch normalization, and ReLU activation function with convolution kernel 1, stride equaling 1, padding equaling 1, and $\text{Concat}(\cdot)$ stands for concatenating the feature maps.

It is proved that this module achieves a good performance on extracting the representative information from the feature map through suitable operation on the information and comprehensively fuse the features. In other words, it combines the feature information of changed and unchanged areas and focuses on the information with higher representative meaning.

B. High-Level Information Refinement Module

After extracting the features in the primary stage, we consider extracting features with high-level information, which refers to abstract, complex, and semantic information extracted from the data, often involving the relationships, context, and meaning associated with the objects or patterns present. Compared with other common methods, our proposed module can obtain the feature map precisely, in an adaptive way. High-level information is usually derived from low-level features, which represent more basic and local patterns or structures in the data, such as edges, colors, and textures. From the aspect of change detection in remote sensing images, high-level feature information involves understanding the complex and abstract patterns related to the changes occurring in the landscape over time, which is extracted from low-level features detected in the initial layers of a deep learning model, such as a CNN. In our research, a deep learning model is trained to detect deforestation using pairs of satellite images taken at different time periods (T1 and T2). More specifically, in the scenario of change detection in remote sensing images, high-level information could include the following.

1) *Land cover classification:* Discerning specific land cover categories, such as forests, agricultural lands, urban areas, or water bodies, by combining low-level features, such as color and texture.
2) *Change patterns:* Comprehending patterns of change between T1 and T2 images, including deforestation (the conversion of forested areas to nonforested land), reforestation (the transformation of nonforest areas into forest), or urban expansion (the development of nonurban land into urban zones).
3) *Contextual information:* Interpreting the spatial relationships and context surrounding the observed changes, such as deforestation events occurring near roads, rivers, or urban locales, which can offer insights into the underlying driving forces.
4) *Temporal information:* Inferring the temporal characteristics of landscape changes, such as the deforestation rate, seasonality, or the regeneration of vegetation following a disturbance.

In our research, which is based on the scenario of change detection in remote sensing imagery, high-level information serves a critical role in understanding the intricate patterns of landscape changes, including deforestation, and enables the proposed deep learning model to accurately detect and interpret changes in remote sensing images across time. Hence, we propose a module called HIRM, which is shown in Fig. 1, in the orange box surrounded by red dashed line. In this module, we also design a submodule called GFFM to fuse four feature maps in different sizes and then restore them back into four different sizes, where we can get four feature maps, which can be denoted as $f_{(Pi(i=3,4,5))}$. Then, we conduct feature refinement through using these feature maps by first concatenating them with the outcome feature map from PFFM, respectively, and send the feature map into a 3-by-3 convolution block, which can be denoted as

$$f_{(Ci(i=3,4,5))} = f^{3\times3}(\text{Concat}(f_{i(i=2,3,4,5)}, f_{Pi(i=3,4,5)})) \tag{9}$$

where $f_{i(i=2,3,4,5)}$ represents for processed four feature map in different scales from PFFM. Then, we refine the high-level semantic information through add the feature map from module PFFM and send the feature map into a 3×3 convolution block. The result feature map is also added with the output feature map $f_{Ri(i=3,4,5)}$ from the previous convolution block. The process can be represented as

$$f_{Ri(i=3,4,5)} = f^{3\times3}(f^{3\times3}(f_{Ci(i=3,4,5)} + f_{i(i=2,3,4,5)}) + f_{Ci(i=3,4,5)}) \tag{10}$$

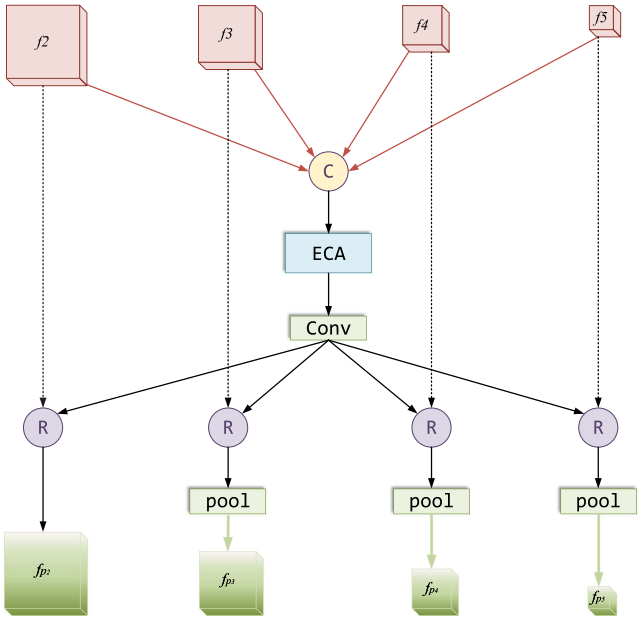

Fig. 5. Global feature fusion module (GFFM).

where $f_{Ci(i=3,4,5)}$ stands for the outcome feature map of the previous formula, and $f_{i(i=2,3,4,5)}$ represents for processed four feature map in different scales from PFFM.

1) Global Feature Fusion Module (GFFM): Feature fusion in CNNs is a technique that combines information from different layers or sources to improve the overall performance of the model. In the context of change detection in remote sensing, feature fusion can help to identify and monitor changes in land cover, urban growth, or natural disasters more accurately. In remote sensing, change detection is the process of identifying differences between satellite or aerial images taken at different times. This task can be challenging due to various factors, such as variations in illumination, atmospheric conditions, and sensor noise. Feature fusion can alleviate these issues by combining information from multiple sources or levels of abstraction in CNN. There are various approaches to implementing feature fusion, given in the following.

1) *Skip connections:* Introduce skip connections between different layers in the CNN, allowing the model to learn multiscale features and capture changes at various resolutions.
2) *Multilevel feature fusion:* Merge features from different layers in the CNN to create a more robust feature representation. This can be done using concatenation or elementwise summation.
3) *Multisource fusion:* Combine information from different sources, such as optical and radar images or different spectral bands, to improve the model's performance.

In our proposed module GFFM, which is displayed in Fig. 5, we use multilevel feature fusion, that is, to merge feature maps from different layers. Before entering GFFM, in the previous module PFFM, we design to upsample one of the outputs, that is, a2, in order to uniform the size of each feature map. Thus, aimed at fusing the features from different scales, we concatenate the four feature maps together. The concatenated feature map can be represented as f_c

$$f_c = \text{Concat}(f_2, f_3, f_4, f_5) \tag{11}$$

where $f_2, f_3, f_4,$ and f_5 stand for four feature maps from different scales. Actually, in remote sensing images, after passing through several convolutional layers, there are often a large number of feature channels, representing different spectral bands, textures, and other image features. However, not all of these features are equally important for detecting changes between images. Moreover, in remote sensing applications, the processing of large-scale images can be computationally intensive, especially if the CNN needs to analyze multiple images at once. To solve these two issues, we add an efficient channel attention (ECA) block. The ECA block can improve feature relevance by helping the CNN to learn to selectively enhance or suppress the most relevant feature channels, leading to better detection performance. In addition, ECA is designed to be computationally efficient, using a 1-D convolutional filter to compute the attention weights for the feature channels. This can help to reduce the overall computational burden and speed up the change detection process.

Hence, after concatenating four feature maps from different scales, we send the concatenated feature map into the ECA block, then the result feature map passes through an ECA block, with a 1-D convolution block where kernel size is 3, where the result can be represented as f_e

$$f_e = \text{ECA}(f_c). \tag{12}$$

In the formula, $\text{ECA}(\cdot)$ denotes the ECA block, where there is a kernel size equaling 3 and padding equaling 1.

Then, fuse the feature map by passing it through a 2-D convolutional block. The fused feature map is refined by a refinement block, which implements through adding two relative feature map and send it to a 2-D convolutional layer with kernel size equaling to 3. This module takes into account the previous feature maps at different scales and the fused feature maps, ensuring the preservation of important features of the original feature maps at various scales, while at the same time taking into account the impact of the global features. Considering restoring the processed feature, an adaptive pooling block is added, resulting in an outcome of four feature maps in four different scales, which can be denoted as $f_{P2}, f_{P3}, f_{P4},$ and f_{P5}, representing the outcome from four branches Res-2, Res-3, Res-4, Res-5, respectively

$$f_{Pi(i=2)} = f^{3\times3}(f_2 + f^{3\times3}(f_e)) \tag{13}$$

$$f_{Pi(i=3,4,5)} = \text{AvgPool}(f^{3\times3}(f_{i(i=3,4,5)} + f^{3\times3}(f_e))) \tag{14}$$

where f_e is the output feature map of the ECA block, and $\text{AvgPool}(\cdot)$ stands for 2-D average pooling layer, where the pooling kernel size and stride are varied. When $i = 2$, there is no average pooling because the major aim of the pooling layer is to achieve feature map restoration, namely, to recover the size of the feature map to its original size. When $i = 3$, the kernel size and stride are 2. When $i = 4$, the kernel size and stride are 4. When $i = 5$, the kernel size and stride are 8. The size of each

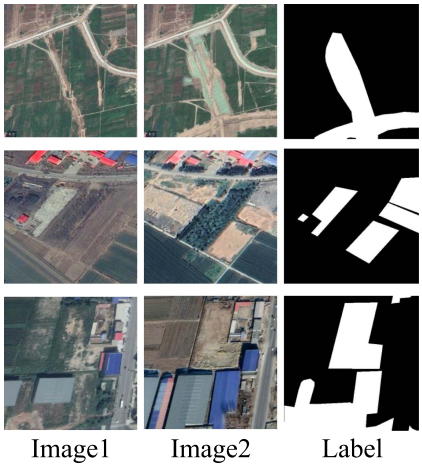

Fig. 6. Partial sample diagrams of LEVIR-CD.

Fig. 7. Partial sample diagrams of CDD.

output can be calculated through the formula

$$\text{out}(N_i, C_i, h, w) = \frac{1}{\text{kH} \cdot \text{kW}} \sum_{m=0}^{\text{kH}-1} \sum_{n=0}^{\text{kW}-1} \text{input}(N_i, C_i \text{stride}[0] \times h + m, \text{stride}[1] \times w + n) \tag{15}$$

where given the input feature map size is input (N,C,H,W), and the output feature map size is out(N_i, C_i, h, w), and the pooling kernel size is (kH, kW).

III. Datasets

We evaluate our proposed SASiamNet using the publicly available datasets LEVIR-CD [25] and CDD [33] because a dataset that is sufficiently sizable can be persuasive in demonstrating the model's capability in handling remote sensing change detection tasks. The performance of the model in complicated urban situations is verified by LEVIR-CD, which is partially depicted in Fig. 6; the performance of the model in dealing with CD tasks in various seasons is verified by CDD, which is partially depicted in Fig. 7.

A. LEVIR-CD Dataset

The LEVIR-CD dataset is a vast collection of high-resolution images that is designed to detect changes in buildings. Comprised of 637 pairs of images obtained from Google Earth, the dataset is focused on monitoring significant land changes that have occurred over a period of 5–14 years, with a particular emphasis on changes in buildings. The dataset encompasses various types of buildings, including villas, high-rise apartments, warehouses, garages, and other similar facilities. Moreover, the dataset takes into account environmental factors, such as light conditions and seasonal variations, which can play a significant role in detecting changes in buildings over time. Overall, the LEVIR-CD dataset offers a valuable resource for researchers and practitioners working in the field of remote sensing, facilitating the development and evaluation of algorithms for detecting and monitoring changes in urban landscapes.

B. CDD Dataset

The CDD dataset is a frequently utilized publicly accessible change detection dataset that takes into consideration variations caused by seasonal changes. The dataset includes 11 pairs of bitemporal remote sensing images with a spatial resolution ranging from $3 - 100$ cm/pixel. Seven of the pairs have a resolution of 4725×2700, while the remaining four pairs have a resolution of 1900×1000. The dataset encompasses various change categories, such as cars, roads, and different large buildings. Due to GPU device limitations, all image pairs have been cropped to a size of 256×256 pixels to establish the training, validation, and test sets, which are comprised of 10 000, 3000, and 3000 images, respectively, as shown in Fig. 8.

IV. Experimental Analysis

All related experiments in our research are conducted on the GeForce RTX 3090 and are based on PyTorch. In addition, the Adam optimizer and BCEWithLogitsLoss loss function are utilized in the neural network. During the network training phase, the learning rate is identified as a crucial hyperparameter. To ensure optimal performance, the ploy method is employed to dynamically adjust the learning rate, initially set to 0.0015. The decay index is set to 0.9, while the batch size and maximum training iterations were configured as 16 and 200, respectively.

Ablation and comparing experiments are conducted on the CDD and LEVIR-CD datasets, respectively. The results revealed that our proposed algorithm outperformed other methods in change detection. To quantify the performance, PR, recall (RC), mean intersection over union (MIoU), and pixel accuracy (PA) were used as the quantitative indicators. The MIoU of the change and nonchange categories were used as the primary evaluation metrics. The calculation formula is shown as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

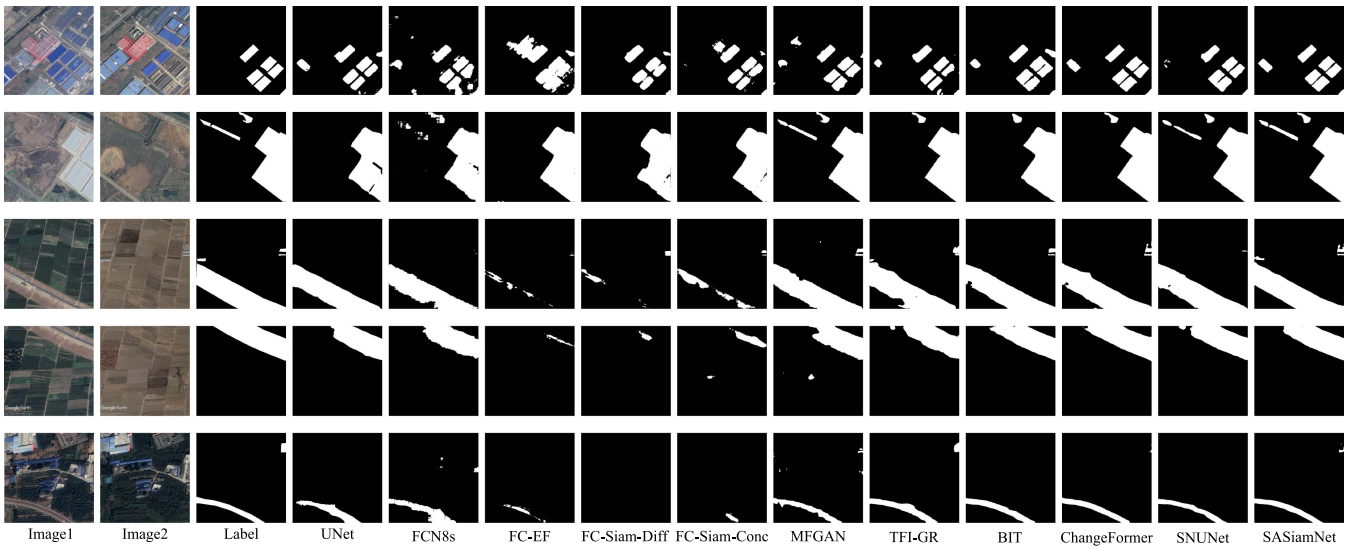

Fig. 8. Comparison diagrams of experiment results on LEVIR-CD.

TABLE I
RESULTS FROM USING RESNET WITH DIFFERENT DEPTHS ON LEVIR-CD

Method	PA (%)	RC (%)	PR (%)	MIoU (%)
SASiamNet_34	94.63	73.24	78.92	82.25
SASiamNet_50	95.17	72.89	79.98	82.99
SASiamNet_18	**97.68**	**75.90**	**83.02**	**85.13**

The bold numbers denote the optimal results.

TABLE II
ABLATION RESULTS ON TWO AUXILIARY MODULES: PFFM AND HIRM

Method	PA(%)	RC(%)	PR(%)	MIoU(%)
ResNet18	93.86	61.45	73.38	75.91
ResNet18+PFFM	95.62	68.71	82.81	81.81
ResNet18+HIRM	91.26	45.36	64.72	68.95
ResNet18+PFFM+HIRM	**97.68**	**75.90**	**83.02**	**85.13**

Bold entities indicate the best score.

$$\text{MIoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{18}$$

$$\text{PA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{19}$$

In these formulas, TP stands for true positive, referring to change area that is correctly predicted, FP stands for false positive, referring to the unchanged area is incorrectly predicted as the part of the changed area, TN stands for true negative, referring to the unchanged area that is correctly predicted, and FN stands for false negative, referring to the part that incorrectly predicts the changed area as unchanged.

A. Network Backbone Selection

We first test the effectiveness of our suggested network by altering the depth of neural networks while maintaining the default values for every training parameters. This was done in order to pick the best backbone network. The tests were conducted using ResNet18, ResNet34, and ResNet50, and the results are given in Table I. ResNet18, ResNet34, and ResNet50 from the LEVIR-CD dataset are represented by SASiamNet_18, SASiamNet_34, and SASiamNet_50 in the table. It has been demonstrated that our suggested network, which uses ResNet18 as its core, performs optimally for change detection in remote sensing images.

B. Ablation Experiments

To assess how well each module performs, we conduct ablation experiments on the LEVIR-CD dataset. When creating complicated neural networks, ablation research is crucial since it increases the effectiveness of network development while also allowing for a deeper knowledge of the proposed network. To test each module's efficacy, we specifically changed the network in this experiment by removing or adding the suggested blocks in the backbone network. In order to evaluate each module's performance, the assessment mostly employed the MIoU statistic. The results of the ablation trials are given in Table II, where the training method for every module remained consistent with the planned network. These results prove the planned network performed superbly, according to the research.

1) Ablation Experiment of PFFM: The PFFM can gain the feature at an early stage, especially the difference and similarities between two images in the same region, and has the capacity of resisting the different environments and all factors that could influence the result of change detection of remote sensing with a slight deviation. It is proved that with the PFFM, we have a better performance on the LEVIR-CD. Namely, the numerical results in Table II reveal that PFFM increases the MIoU score by 5.90%.

2) Ablation Experiment of HIRM: The HIRM aggregates and refines global information across different scales, focusing on high-level features for more precise semantic information. It adaptively extracts high-level features from multiple scales,

TABLE III
ABLATION EXPERIMENT OF PFFM THROUGH ADDING SUBMODULES IN PFFM, WHERE PFFM_0 STANDS FOR THE BLOCK PFFM WITHOUT ANY SUBMODULE ADDED, AND PFFM_{AM} STANDS FOR THE BLOCK PFFM WITH ONLY ONE SUBMODULE AM, AND VICE VERSA

Method	PA(%)	RC(%)	PR(%)	MIoU(%)
PFFM_0	94.29	63.03	74.64	79.32
PFFM_{AM}	93.19	68.73	77.84	82.17
PFFM_{DM}	94.36	71.74	76.35	82.01
$\text{PFFM}_{\text{RFEM}}$	96.52	73.40	76.22	82.98
$\text{PFFM}_{\text{AM+DM}}$	95.78	74.07	82.18	83.47
$\text{PFFM}_{\text{AM+RFEM}}$	95.71	71.52	80.59	82.97
$\text{PFFM}_{\text{DM+RFEM}}$	96.32	75.20	82.72	84.01
$\text{PFFM}_{\text{AM+DM+RFEM}}$	**97.68**	**75.90**	**83.02**	**85.13**

Bold entities indicate the best score.

TABLE IV
ABLATION EXPERIMENT OF HIRM THROUGH CHANGING THE NUMBER OF HIRM

Number of HIRM	PA(%)	RC(%)	PR(%)	MIoU(%)
1	95.71	75.23	80.03	81.49
3	94.53	74.48	77.82	83.21
4	**97.77**	74.98	79.34	81.43
2	97.68	**75.90**	**83.02**	**85.13**

Bold entities indicate the best score.

amalgamating significant features into a comprehensive feature map. This process, enhancing meaningful and suppressing irrelevant features, results in improved prediction accuracy. As seen in Table II, adding HIRM elevates the MIoU score by 3.32% compared with ResNet18+PFFM, indicating superior semantic information extraction. However, it is effective only in conjunction with PFFM, suggesting their interdependent relationship in network performance enhancement. Therefore, primary stage feature extraction is a prerequisite for effective information refinement.

1) *Ablation experiment of PFFM:* During the design of the PFFM, we incorporated three submodules: the DM, AM, and RFEM. To evaluate the effectiveness of these submodules, we conducted an ablation comparison experiment on the PFFM. Table III presents the results, where PFFM_0 refers to the PFFM block without any submodule, PFFM_{AM} represents the PFFM block with only the AM submodule, and vice versa.
   Furthermore, we also conduct an ablation experiment on the count of HIRM. As given in Table IV, two layers of HIRM can effectively aggregate and refine the high-level image information, while too many layers may cause issues, such as overfitting.
   The findings demonstrate that the inclusion of the AM, DM, and RFEM submodules leads to improvements in the MIoU score. Specifically, the MIoU score increases by 2.85%, 2.69%, and 3.66% with the addition of AM, DM, and RFEM, respectively. Furthermore, when AM and DM work together to detect changed and unchanged areas in different periods, the experimental results show a combined improvement of 4.15%. This indicates that these two modules complement each other's functions effectively. Moreover, the addition of RFEM, which selectively extracts the most representative features from the feature map, further enhances the MIoU score by 1.66% compared with the results obtained by only adding AM and DM to the PFFM. These experimental findings demonstrate the collaborative and supportive nature of these three submodules, leading to enhanced performance of our network.

TABLE V
ABLATION EXPERIMENT OF HIRM THROUGH ADDING THE SUBMODULE GRRM IN HIRM AND CHANGING THE CONNECTION OPERATION IN GFFM

Method	PA(%)	RC(%)	PR(%)	MIoU(%)
HIRM_0	94.16	65.25	74.87	80.12
$\text{HIRM}_{\text{GFFM}_+}$	94.93	74.82	77.45	82.76
$\text{HIRM}_{\text{GFFM}_c}$	**97.68**	**75.90**	**83.02**	**85.13**

Bold entities indicate the best score.

2) *Ablation experiment of HIRM:* Aimed at verifying the effectiveness of HIRM, we also carry out two ablation experiments on HIRM. First, we set the number of block HIRM as 1, and we then conduct the ablation experiment on the submodule of HIRM, that is, GFFM. Table V gives the result of the experiment, where HIRM_0 stands for block HIRM without submodule, $\text{HIRM}_{\text{GFFM}_+}$ stands for block HIRM with submodule GFFM, whose fusion approach for aggregating the feature map in four scales is to add them together, and $\text{HIRM}_{\text{GFFM}_c}$ stands for block HIRM with submodule GFFM, whose fusion approach is concatenating. It is proved that with the insert of GFFM, where concatenating is applied, can achieve a better score of MIoU. In Table V, the MIoU scores of our proposed model are increased by 4.13% and 1.34%, respectively, compared with no submodule GFFM within HIRM, and elementwise summation in GFFM, respectively.

C. Comparative Experiments

To provide a comprehensive evaluation of our proposed model, we conduct a comparative analysis using two open-source datasets, assessing its performance in diverse change detection scenarios, including different lighting conditions, complex urban buildings, and wasteland areas. We compare our method with various traditional unsupervised algorithms (IR-MAD and PCA-Means), semantic segmentation algorithms based on CNN (UNet, DeepLabv3+ and BiseNet), semantic segmentation algorithms based on Transformer (ViT and PvT), CNN-based change detection algorithms (FC-EF, FC-CONC, FC-DIFF, TFI-GR, SNUnet, DTCDSCN, STANet and IFNet), and hybrid CNN and Transformer-based change detection algorithms (BIT, TransUNetCD, UVACD and TransCD). All deep learning models in the comparison experiment undergo the same training modes to ensure fairness and objectivity.

Evaluation results for the LEVIR-CD and CDD datasets are given in Table VI. Traditional unsupervised methods, such as IR-MAD and PCA-Means, have low accuracy and struggle with identifying change regions. While CNN-based change detection methods generally perform better, they exhibit lower accuracy than hybrid CNN and Transformer-based models due to limitations in capturing global information. Despite their ability to

TABLE VI
RESULTS OF COMPARISON EXPERIMENTS ON TWO PUBLIC DATASETS, LEVIR AND CDD

Method	Metrics	
	(LEVIR)	(CDD)
	PA (%) / RC (%) / PR (%) / MIoU (%)	PA (%) / RC (%) / PR (%) / MIoU (%)
IR-MAD [34]	27.52 / 41.35 / 13.58 / 31.24	23.91 / 35.82 / 12.79 / 25.63
PCA-Means [14]	31.24 / 46.29 / 17.81 / 37.12	29.31 / 38.54 / 15.87 / 29.75
UNet [29]	96.77 / 70.25 / 71.15 / 78.97	94.03 / 45.81 / 56.79 / 62.58
ViT [35]	96.40 / 70.54 / 70.75 / 77.16	94.70 / 47.35 / 58.09 / 66.92
DeepLabV3+ [30]	96.13 / 68.92 / 70.42 / 77.81	95.87 / 51.52 / 62.11 / 70.30
BiseNet [37]	95.03 / 67.17 / 68.79 / 74.82	96.03 / 52.19 / 61.40 / 69.58
FC-EF [24]	95.25 / 68.72 / 70.35 / 76.86	95.58 / 50.82 / 61.51 / 68.73
FC-Siam-Diff [24]	91.20 / 42.99 / 71.83 / 66.82	94.00 / 25.78 / 33.94 / 55.38
FC-Siam-Conc [24]	91.11 / 49.00 / 69.93 / 68.30	93.62 / 20.59 / 31.23 / 53.74
FCN-8s [31]	92.05 / 60.50 / 65.34 / 71.92	94.67 / 41.09 / 48.54 / 61.85
PvT [39] 37	95.97 / 70.22 / 71.34 / 78.41	95.19 / 51.27 / 60.61 / 68.33
DTCDSCN [38]	95.83 / 69.27 / 71.37 / 78.05	96.45 / 55.93 / 69.81 / 67.62
STANet [25]	95.37 / 70.14 / 71.05 / 78.27	96.02 / 53.41 / 62.54 / 71.07
IFNet [40]	96.16 / 70.32 / 71.87 / 79.58	96.46 / 54.97 / 63.58 / 71.96
TransCD [41]	96.84 / 71.20 / 71.42 / 79.34	95.37 / 55.73 / 63.81 / 71.60
ChangeFormer [45]	96.19 / 72.56 / 82.34 / 83.21	96.58 / 64.71 / 60.59 / 71.28
MFGAN [32]	94.10 / 67.27 / 74.16 / 77.76	97.22 / 62.69 / 56.96 / 71.14
BIT [26]	97.35 / 71.62 / 72.54 / 81.06	96.03 / 55.95 / 63.41 / 72.53
UVACD [41]	96.97 / 70.65 / 71.97 / 81.23	95.70 / 55.28 / 62.94 / 72.06
TransUNetCD [42]	97.08 / 71.73 / 71.83 / 81.70	96.41 / 55.52 / 63.27 / 72.89
TFI-GR [43]	95.97 / 74.97 / 81.70 / 84.02	97.32 / 68.39 / 54.10 / 72.58
SNUNet [44]	97.03 / 70.71 / 72.25 / 80.47	96.18 / 56.03 / **64.28** / 73.09
SASiamNet (ours)	**97.68** / **75.90** / **83.02** / **85.13**	**97.76** / **69.02** / 64.19 / **73.97**

Bold represents the best result.

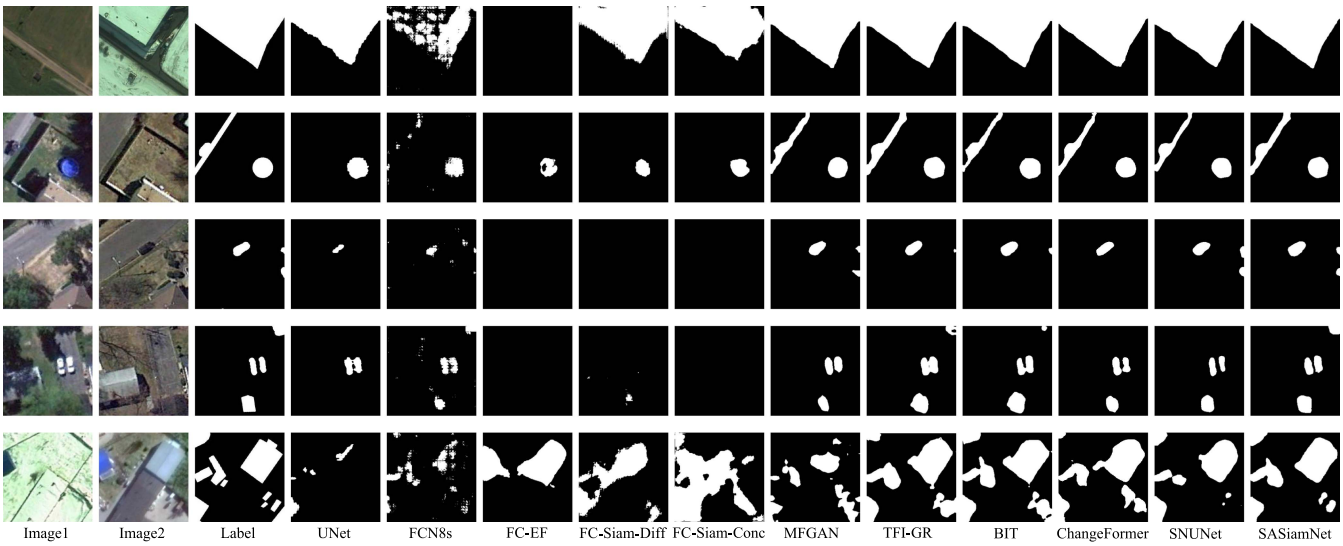

Fig. 9. Comparison diagrams of experiment results on CDD.

capture global context, Transformer-based models may neglect critical local features. To address this, we propose SASiamNet, a Siamese network with a U-shaped backbone structure. The PFFM ensures adequate multilevel feature information extraction, further refined by the HIRM. On the LEVIR and CDD datasets, SASiamNet achieves PA indices of 97.68 and 97.76, RC indices of 75.90 and 69.02, PR indices of 83.02 and 64.19, and MIoU indices of 85.13 and 73.97, respectively, validating its robustness for change detection in high-resolution remote sensing. SASiamNet consistently outperforms comparative algorithms.

Fig. 9 shows the prediction outcomes from different techniques on two public datasets. Traditional unsupervised methods struggle to identify changing areas, and semantic segmentation algorithms generate many false detections in complex scenes. Conversely, change detection algorithms enhance overall predic-

tion accuracy by accurately recognizing change areas, reducing false detections, and limiting missed detections. However, these techniques face issues when dealing with pixel discrepancies due to seasonal changes or inconsistent sensor angles, primarily due to subpar location and semantic information extraction and fusion. As shown in Fig. 10, comparison algorithms produce rough edge predictions in these scenarios. However, our proposed SASiamNet method, using a U-shaped backbone network with a Siamese structure and PFFM, effectively overcomes these limitations by extracting and learning multilevel feature information. The superior performance of SASiamNet in complex change detection tasks is further validated by results from three public datasets.

V. Conclusion

The present study introduces a SASiamNet designed to discern changes in high-resolution remote sensing imagery. This proposed network showcases outstanding performance in real-time land cover segmentation tasks. It employs ResNet as the primary architecture, which is used to extract both local and global information from high-resolution remote sensing images. The network incorporates the PFFM, responsible for the extraction and fusion of primary stage feature maps. This is followed by the HIRM, which refines the extracted feature map and efficaciously transmutes low-level semantic information into high-level semantic information. The efficacy of SASiamNet is evaluated using two datasets, LEVIR-CD and CDD. These datasets comprise bitemporal images sourced from Google Earth, covering various regions across China. Experimental results illustrate that the proposed technique surpasses traditional methodologies and current leading-edge change detection methods in performance. Nevertheless, there remains an opportunity for further improvement in the proposed algorithm, primarily due to the intricate nature of the model. As a result, subsequent research must focus on reducing model complexity while preserving the accuracy of detection results. Future work will aim to decrease the number of parameters without adversely affecting the model's change detection accuracy.

Acknowledgment

CRediT authorship contribution statement: Xianxuan Long—Conceptualization, Methodology, Writing—original draft. Wei Zhuang—Conceptualization, Methodology, Writing—review and editing, Supervision, Project administration. Min Xia—Conceptualization, Supervision, Software, Writing—review and editing, Funding acquisition. Kai Hu—Supervision, Writing—review and editing. Haifeng Lin—Validation, Writing—review and editing.

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

References

[1] C. Cleve, M. Kelly, F. R. Kearns, and M. Moritz, "Classification of the wildland–urban interface: A comparison of pixel-and object-based classifications using high-resolution aerial photography," *Comput., Environ. Urban Syst.*, vol. 32, pp. 317–326, 2008.

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

[3] C. Ma, H. Yin, L. Weng, M. Xia, and H. Lin, "DAFNet: A novel change-detection model for high-resolution remote sensing imagery based on feature difference and attention mechanism," *Remote Sens.*, vol. 15, 2023, Art. no. 3896.

[4] K. Chen, X. Dai, M. Xia, L. Weng, K. Hu, and H. Lin, "MSFANet: Multi-scale strip feature attention network for cloud and cloud shadow segmentation," *Remote Sens.*, vol. 15, 2023, Art. no. 4853.

[5] X. Dai, K. Chen, M. Xia, L. Weng, and H. Lin, "LPMSNet: Location pooling multi-scale network for cloud and cloud shadow segmentation," *Remote Sens.*, vol. 15, 2023, Art. no. 4005.

[6] L. I. Kuncheva and W. J. Faithfull, "PCA feature extraction for change detection in multidimensional unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 1, pp. 69–80, Jan. 2014.

[7] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering," *IEEE Geosc. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[8] A. Singh and K. K. Singh, "Unsupervised change detection in remote sensing images using fusion of spectral and statistical indices," *Egyptian J. Remote Sens. Space Sci.*, vol. 21, pp. 345–351, 2018.

[9] J. Deng, K. Wang, Y. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, pp. 4823–4838, 2008.

[10] S. Juan, W. Gui-Jin, L. Xing-Gang, and L. Dai-Zhi, "A change detection algorithm for man-made objects based on multi-temporal remote sensing images," *Acta Automatica Sinica*, vol. 34, pp. 1040–1046, 2008.

[11] C. Lu, M. Xia, M. Qian, and B. Chen, "Dual-branch network for cloud and cloud shadow segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410012.

[12] J. Chen, M. Xia, D. Wang, and H. Lin, "Double branch parallel network for segmentation of buildings and waters in remote sensing images," *Remote Sens.*, vol. 15, 2023, Art. no. 1536.

[13] C. Zhang, L. Weng, D. Li, M. Xia, and H. Lin, "CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery," *Remote Sens.*, vol. 15, 2023, Art. no. 1664.

[14] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[15] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Inf. Sci.*, vol. 535, pp. 156–171, 2020.

[16] L. Weng, K. Pang, M. Xia, H. Lin, M. Qian, and C. Zhu, "Sgformer: A local and global features coupling network for semantic segmentation of land cover," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6812–6824, 2023.

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

[18] D. Wang, L. Weng, M. Xia, and H. Lin, "MBCNet: Multi-branch collaborative change-detection network based on Siamese structure," *Remote Sens.*, vol. 15, 2023, Art. no. 2237.

[19] H. Ji, M. Xia, D. Zhang, and H. Lin, "Multi-supervised feature fusion attention network for clouds and shadows detection," *ISPRS Int. J. Geo-Inf.*, vol. 12, 2023, Art. no. 247.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[21] 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.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[23] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "DenseNet: Implementing efficient ConvNet descriptor pyramids," 2014, *arXiv:1404.1869*.

[24] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[25] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, 2020, Art. no. 1662.
[26] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[27] Q. Ding, Z. Shao, X. Huang, and O. Altan, "DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, 2021, Art. no. 102591.
[28] X. Wang et al., "A high-resolution feature difference attention network for the application of building change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102950.
[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.
[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[31] 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.
[32] S. Chu, P. Li, and M. Xia, "MFGAN: Multi feature guided aggregation network for remote sensing image," *Neural Comput. Appl.*, vol. 34, pp. 10 157–10 173, 2022.
[33] 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.
[34] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.
[35] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[36] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.
[37] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, 2021, Art. no. 102348.
[38] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[39] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.
[40] Z. Wang, Y. Zhang, L. Luo, and N. Wang, "TransCD: Scene change detection via transformer-based architecture," *Opt. Exp.*, vol. 29, no. 25, pp. 41 409–41 427, 2021.
[41] G. Wang, B. Li, T. Zhang, and S. Zhang, "A network combining a transformer and a convolutional neural network for remote sensing image change detection," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2228.
[42] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.
[43] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628711.
[44] 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, 2022, Art. no. 8007805.
[45] W. Bandara and V. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.