U-Shaped CNN-ViT Siamese Network With Learnable Mask Guidance for Remote Sensing Building Change Detection

Yongjing Cui, *Student Member, IEEE*, He Chen, *Member, IEEE*, Shan Dong, *Member, IEEE*, Guanqun Wang, *Member, IEEE*, and Yin Zhuang, *Member, IEEE*

***Abstract*—Building change detection (BCD) aims to identify new or disappeared buildings from bitemporal images. However, the varied scales and appearances of buildings, along with the challenge of pseudochange interference from complex backgrounds, make it difficult to accurately extract complete changes. To address these challenges in BCD, a U-shaped hybrid Siamese network combining a convolutional neural network and a vision transformer (CNN-ViT) with learnable mask guidance, called U-Conformer, is designed. First a new hybrid architecture of U-Conformer is proposed. The architecture integrates the strengths of CNNs and ViTs to establish a robust, multiscale heterogeneous representation that aids in detecting buildings of various sizes. Second, a learnable mask guidance module is specifically designed for U-Conformer, focusing the multiscale heterogeneous representation on extracting relevant scale changes while progressively suppressing pseudochanges. Furthermore, for the U-Conformer architecture, a mask information joint class-balanced loss function that combines the binary cross-entropy loss function and the dice loss function is devised, significantly mitigating the issue of class imbalance. Experimental results on three publicly available change detection datasets, LEVIR-CD, WHU-CD, and GZ-CD, demonstrate that U-Conformer surpasses previous methods, achieving F1 scores of 91.5%, 94.6%, and 86.7%, as well as IoU scores of 84.3%, 89.7%, and 76.5% on the LEVIR-CD, WHU-CD, and GZ-CD datasets, respectively.**

***Index Terms*—Building change detection (BCD), learnable mask guidance, multiscale representation, remote sensing, U-shaped convolutional-neural-network-vision-transformer (CNN-ViT).**

I. Introduction

BUILDING change detection (BCD) has to identify and analyze building change information from complex remote sensing scenes within a specific area over time, such as construction, demolition, or renovations [1], [2]. This technology is frequently used in urban development planning [3], [4], monitoring environmental impacts [5], [6], assessing the consequences of natural disasters [7], [8], and aiding decision making in land use and policy formulation [9], [10]. BCD utilizes various data sources, including satellite image and aerial photography. In remote sensing images, different buildings have various scales and appearances, resulting in notable intraclass differences, as illustrated in Fig. 1. Remote sensing images often feature both large-scale buildings, such as hospitals and factories, as shown in Fig. 1(c), and small-scale structures like houses and farms, as shown in Fig. 1(b). Consequently, we have analyzed the range of building area changes in three datasets, as illustrated in Fig. 2(d). We find that the scale variation of buildings in remote sensing images is drastic, demanding high capacity for multiscale perception from the networks. In addition, variations in building density [from densely populated communities to sparsely distributed villas, as depicted in Fig. 1(a) and (b)] compound the complexity of the backgrounds against which building targets are positioned. Finally, as seen in Fig. 1(b), remote sensing images also include special buildings with unique morphological structures (e.g., supermarkets and churches), characterized by rich and varied appearances. Moreover, the presence of pseudochanges induced

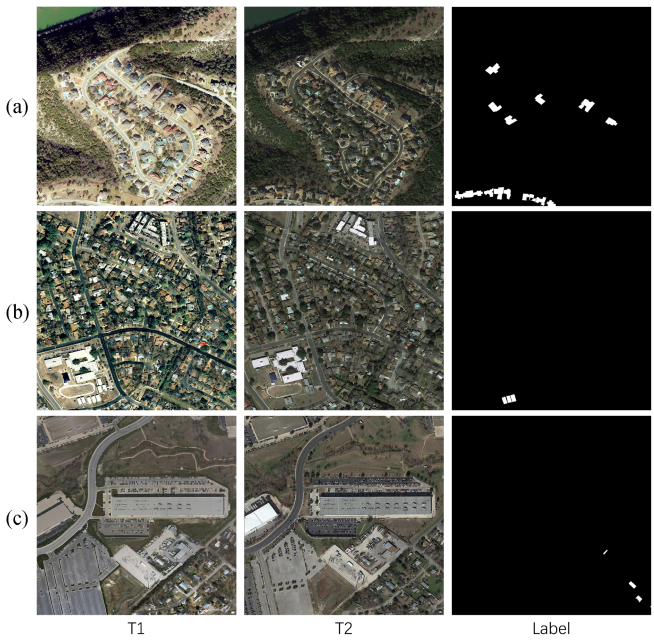

Fig. 1. Examples of the LEVIR-CD dataset. T1 represents T1-time image, T2 represents T2-time image, and Label represents the GT map.

Manuscript received 18 January 2024; revised 23 April 2024; accepted 28 May 2024. Date of publication 3 June 2024; date of current version 19 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62371048 and in part by the National Science Foundation for Young Scientists of China under Grant 62101046. *(Corresponding authors: He Chen; Yin Zhuang.)*

The authors are with the National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing, Beijing Institute of Technology, Beijing 100081, China (e-mail: chenhe@bit.edu.cn; yzhuang@bit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3408604

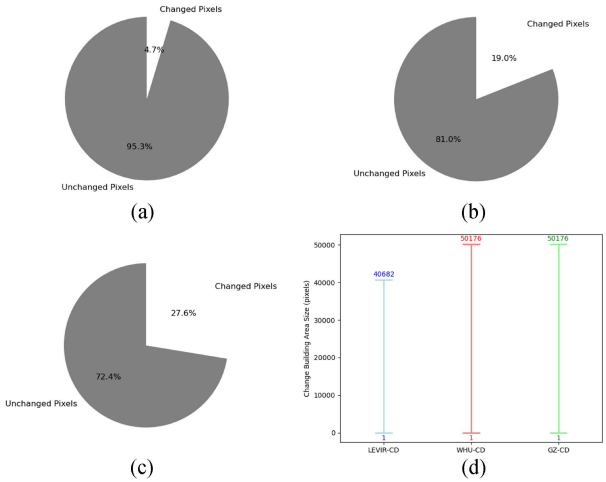

Fig. 2. Ratios of changed and unchanged pixels in three BCD datasets. (a) LEVIR-CD. (b) WHU-CD. (c) GZ-CD. (d) Statistical overview of the scale range of changed areas in these datasets.

by factors such as variations in lighting conditions, weather conditions (e.g., rain, snow, and so on), and imaging equipment settings further complicates the task. Therefore, the key to achieving high-precision BCD lies in effectively decoding all relevant change information from bitemporal remote sensing images to establish efficient representations of building change targets, while simultaneously suppressing or even eliminating interference caused by various pseudochanges. Furthermore, in BCD tasks, there is a widespread issue of class imbalance, as illustrated in Fig. 2(a)–(c). We have analyzed the proportion of changed versus unchanged areas in three public datasets. It is evident that the number of unchanged samples significantly surpasses that of changed samples. This imbalance poses a substantial challenge to BCD models in effectively extracting discriminative features.

Traditional remote sensing image change detection (CD) methods mainly include algebraic methods such as image differencing [11] and image quantification [12], transformation-based methods like change vector analysis [13], [14], principal component analysis [15], [16], and independent component analysis [17], classification-based methods including post-classification methods, and other algorithms like slow feature analysis [18] and multivariate alteration detection [19]. These traditional image CD methods often rely on manual feature design, however, manually designed features lack robustness and struggle with perceiving multiscale targets. Their inadequate capacity to suppress pseudochanges leads to unsatisfactory detection results. Therefore, the field of remote sensing image CD has been seeking breakthroughs.

As deep learning technologies have advanced, they have demonstrated a remarkable ability to autonomously learn and adapt to target information in a data-driven fashion. The autonomously learned features of deep learning have proven to be significantly more robust than traditional, manually crafted features. This realization has ignited a surge of research in CD methods, leveraging the power of deep learning. In 2018, Daudt et al. [20] designed three innovative fully convolutional neural network (CNN) architectures based on the U-Net structure: FC-EF, FC-Siam-Conc, and FC-Siam-Diff. Particularly noteworthy are FC-Siam-Conc and FC-Siam-Diff, which used skip connections to merge early spatial details with later abstract features in Siamese encoding–decoding architectures. These quickly gained widespread acceptance in the BCD field due to their outstanding effects. Following this, research efforts concentrated on enhancing these models' ability to perceive information across multiple scales, aiming for superior feature representation. In 2020, Chen et al. [21] proposed the DASNet, which utilized a dual-attention mechanism to capture long-range dependencies, leading to more distinct feature representations. In addition, to tackle the challenge of sample imbalance, they designed a weighted bilateral contrast loss, fine-tuning the network to focus on change feature pairs. Further advancements came with STANet [22], which divided images into multiscale subregions. It applied CD self-attention mechanisms to model the spatiotemporal relationships in dual-temporal images, capturing spatiotemporal dependencies at varied scales for enhanced feature representations. In 2022, Hang et al. [23] proposed a multiScale progressive segmentation network, using three subnetworks, a scale guidance module, and a position-sensitive module to detect change targets at various scales. In 2023, Lv et al. [24] proposed a land cover CD network based on hierarchical attention feature fusion, which explores the global semantic features of interesting targets from multiple perspectives through especially designed multiscale convolution fusion filters, achieving detection of multiscale targets. In the same year, Feng et al. [25] designed DMINet to achieve high-precision detection while effectively minimizing false alarms caused by pseudochange interference. This was accomplished through cross-temporal joint attention blocks. They directed the global feature distribution of each input and encouraged information coupling between intralayer representations. In addition, Zhou et al. [26] proposed CANet, which includes temporal attention modules, context memory and extraction modules, and multiscale fusion modules. These designs allow the network to capture intraimage context more effectively by mining interimage context information, thus improving the utilization of image context. Also in 2023, Lv et al. [27] developed E-UNet, integrating multiscale convolution and polarized self-attention modules for multimodal remote sensing image CD. The most recent innovation came in 2024, with AANet, proposed by Hang et al. [28], using an ambiguity refinement module to assist in extracting difference features and a weight rearrangement module to fuse features of different scales, achieving satisfactory CD in ambiguous areas. These models incorporate various attention mechanisms and multiscale feature capture mechanisms, significantly expanding the receptive fields of CNNs and enhancing their ability to assimilate and process global information.

Since the introduction of Transformers in 2017 [29], their exceptional global modeling capabilities have captivated the research community. This interest has spurred numerous studies exploring the application of Transformer-based architectures in computer vision, with the goal of achieving more precise processing outcomes. In the realm of image analysis, this exploration has led to the development of innovative models such as

vision transformers (ViT) [30], segmentation transformers [31], and SegFormer [32]. These Transformer-based networks have notably enhanced the accuracy of CD tasks. A prominent example from 2022 is the ChangeFormer [33], which innovatively integrated Transformers into BCD tasks. The ChangeFormer capitalized on the Transformer's extensive receptive field and advanced contextual modeling capabilities, surpassing traditional CNNs in precision. This breakthrough represents a significant advancement in BCD, showcasing the potential of Transformer-based models to redefine the landscape of image analysis and CD.

However, networks that solely depend on stacked Transformers are limited in their multiscale perception capabilities, often overly focusing on global modeling and overlooking smaller scale targets and their detailed characteristics. Extensive research [34], [35] has shown that solely relying on this method is ineffective for capturing building information across varying scales. As a result, the integration of Transformer structures with CNN architecture has emerged as a pivotal area of research for detecting targets of different sizes in remote sensing images. Chen et al. [36] have pioneered an innovative method with the bitemporal image transformer (BIT), which skillfully models spatial-temporal contexts. This method uses a semantic tokenizer to compactly organize features extracted by a CNN backbone into a set of tokens, which are then processed by the Transformer encoder to establish connections within a token-based spatiotemporal domain. These tokens, enriched with context from each temporal image, are then mapped back into pixel space to refine the initial features through the Transformer decoder. The final step involves the prediction module delivering pixel-level outcomes by filtering the feature differences through a basic CNN. The effectiveness of the BIT highlights the feasibility and benefits of merging CNNs with Transformers, proposing a viable approach for such technological unions. However, this strategy also faces its own challenges, notably the Transformers' requirement for large, diverse training datasets to prevent overfitting and guarantee generalization. Compiling and annotating these vast datasets is not only laborious but also costly, posing a substantial challenge for the application of Transformer architectures in BCD tasks and marking a critical area for future development in these sophisticated systems.

Taking into account the distinct advantages and limitations of CNNs and Transformers, this study proposes the U-Conformer model, specifically designed for BCD using remote sensing images. This innovative model employs a self-supervised multiscale masked autoencoder pretraining scheme (CFMAE) to maximize the potential of unlabeled remote sensing data. The U-Conformer architecture seamlessly combines the strengths of CNNs and ViTs, offering a powerful solution for BCD challenges in remote sensing scenarios. Within this framework, CNNs excel at capturing the intricate local details of objects, providing a detailed perspective of each element in an image. In contrast, ViTs are adept at forming a comprehensive understanding of global semantic relationships, effectively integrating these detailed insights into a unified narrative. This hybrid architecture addresses the common problem of weak features in traditional BCD networks, thereby facilitating a more dependable detection mechanism. The efficiency of this approach is reflected in significant improvements in detection performance, achieved with considerably less computational overhead and a streamlined model structure. The U-shaped architecture of the U-Conformer plays a crucial role in boosting the network's proficiency in accurately identifying and analyzing targets across different scales. This design promotes a deeper comprehension of variations in building structures, regardless of their dimensions. Moreover, the addition of the learnable mask guidance module (LMGM) is instrumental. It sharpens the network's focus, steering it toward relevant changes. This strategy is focusing on mitigating pseudochange interference, a prevalent challenge in CD tasks. As a result, LMGM markedly enhances the network's ability to discern building change information across various scales in remote sensing images. Furthermore, the mask information joint class-balanced loss function strategically tackles the widespread issue of class imbalance in CD datasets. This recalibration of the learning process ensures the algorithm remains impartial, giving enough attention to the often more critical minority class. The implementation of this balanced approach substantially elevates the model's precision and reliability, making the U-Conformer an exceptionally effective tool for detecting architectural changes in remote sensing images.

The proposed U-Conformer is comprehensively tested on three challenging publicly available CD datasets: LEVIR-CD [21], WHU-CD [37], and GZ-CD [38]. The impressive experimental results demonstrate that U-Conformer achieves optimal detection outcomes in both quantitative metrics and visual performance, outperforming ten state-of-the-art (SOTA) methods in terms of performance, and achieving F1 scores of 91.5%, 94.6%, and 86.7%, as well as IoU scores of 84.3%, 89.7%, and 76.5% on the LEVIR-CD, WHU-CD, and GZ-CD datasets, respectively. Our contributions are as follows.

1) This article proposes an effective BCD algorithm that efficiently integrates the strengths of CNN and ViT, achieving high-precision performance in detecting building changes while effectively suppressing background pseudochange interference.
2) This article designs a U-shaped Siamese architecture. To address the challenge of poor multiscale target perception in network for BCD tasks, a U-shaped multiscale structure is considered as the feature encoder–decoder method. This approach integrates the rich detail information from shallow layers with the abstract semantic information from deep layers, enhancing the model's ability to perceive multiscale targets.
3) This article designs the LMGM. To tackle the significant challenge of pseudochange interference and enhance the network's multiscale perception capabilities, the network integrates the LMGM to guide the focus toward authentic multiscale change information while suppressing pseudochange interference. Furthermore, by employing a class-balanced loss function, it effectively resolves the issues caused by class imbalance, thereby significantly improving the accuracy of target detection.

II. Related Work

A. Change Detection (CD)

As the frontier of intelligent systems [39] continues to expand, the role of CD technology has become increasingly vital in the arena of dynamic object tracking [40], [41], [42], [43]. This innovative technology is indispensable for quick and accurate decision making in the face of sudden environmental changes, such as the unexpected appearance of obstacles, pedestrians, or vehicles [44], [45]. Moreover, it plays a pivotal role in the tracking of moving entities, from individuals within a crowd to vehicles navigating busy streets [39], thereby significantly augmenting intelligent applications in areas like surveillance, traffic management, and public safety. The integration of sophisticated artificial intelligence CD algorithms with SOTA sensor technologies not only elevates the accuracy of these systems but also enhances their capability to adapt to and learn from a myriad of scenarios. In the varied application landscape of intelligent systems, natural scene images and videos serve as the primary data sources, offering a wealth of real-time information that is crucial for the effective operation of these advanced systems in dynamic and complex environments.

CD based on natural scene images is widely used in updating 3-D maps. The targets for detection are changes in navigation landmarks in natural scenes, such as buildings, traffic signs, and other roadside structures (fences, kiosks, etc.). This is vital for updating and maintaining urban maps to ensure the robustness of navigation systems. Alcantarilla et al. [46] proposed a system for structural CD in street view videos, meeting the demands for more frequent and efficient large-scale map updates in autonomous vehicle navigation. Wang et al. [47] proposed a scene CD architecture based on the Transformer to overcome pseudochanges caused by camera movement or environmental changes.

On the other hand, CD based on natural scene video is mainly used in automated video surveillance [48]. For example, Haritaoglu et al. [49] developed a surveillance system that employs CD technology to identify individuals and their carried items in a crowd. Carnegie Mellon University [48], on the other hand, utilized CD technology to track moving targets, such as people, cars, and others. Robert O'Callaghan et al. [50] proposed a robust CD algorithm that can accurately segment person changes in a scene while ignoring the effects of lighting changes.

In natural scenes, the scale of target changes is consistent, and the availability of information from various angles ensures minimal intraclass differences for the same targets. This aspect significantly enhances the performance of deep learning techniques in scene CD. However, in the task of CD in remote sensing images, the detection targets are presented only from a bird's-eye view, characterized by large intraclass differences, small interclass differences, large scale variances, and complex backgrounds. This poses challenges in the field of computer vision. As remote sensing technology continues to advance, the need for CD in remote sensing images has grown significantly, rendering it an area of considerable value and interest in research. Therefore, this article focuses on the research of CD technology in remote sensing images, aiming to design an algorithm capable of accurately identifying changes in building targets of various scales and styles.

B. CNN-ViT Hybrid Structure

In the field of computer vision, CNNs, with their hierarchical structure and capability to gradually extract local features, have long dominated due to their ability to consider both low-level geometric features and high-level semantic features [51], [52], [53]. However, their inability to fully utilize global information in images often leads to classification errors when features are limited or ambiguous, prompting researchers to seek new network architectures. The advent of the ViT demonstrated that Transformers could be applied to visual tasks. Extensive experiments have shown that ViT-based models consistently outperform CNN-based models in object-level visual tasks such as object detection and classification. This can be attributed to the built-in self-attention mechanism in the ViT, which effectively captures long-range data dependencies [54], [55], [56]. However, the use of the ViT demands stronger computational capabilities and longer training cycles, contrasting sharply with the requirements of the CNN. In addition, the breadth of training data plays a crucial role in ViTs' performance, with the combination of large-scale training datasets and data augmentation methods ensuring ViTs' superiority over CNNs [57]. However, in pixel-level visual tasks like image segmentation and CD, CNNs still outperform ViTs due to their convolutional operations' proficiency in extracting local features.

Therefore, leveraging the complementary strengths of both and developing a hybrid architecture (Conformer) combining the CNN and ViT has become a hot topic in computer vision. In the field of scene detection, Li et al. [34] combined the CNN and ViT as the backbone to capture both local and global features, proposing an efficient distracted driving detection method using driver and relevant object cues. In the realm of multimodal fusion, Yu et al. [35] designed an unsupervised hybrid model based on the CNN and ViT for multimodal medical image fusion, combining the advantages of both models to capture global contextual information, enhance learning capabilities, and introduce CNN's inductive bias to improve generalization performance.

In the task of CD in remote sensing images, which requires precise detection of small targets and edge details while also considering large-area target perception, the Conformer is particularly well-suited. It leverages both the CNN's detail capturing and localization capabilities and the ViT's global perception and modeling abilities to address the challenge of detecting changes in multiscale architectural targets in complex remote sensing scenes.

III. Methodology

In this section, the overall architecture of the U-Conformer is introduced, as delineated in Section II-A. This is followed by comprehensive explanations of the U-Conformer detailed in Section II-B, and a description of the LMGM in Section II-C. Finally, Section II-D introduces a loss function tailored to mitigate the issue of sample imbalance.

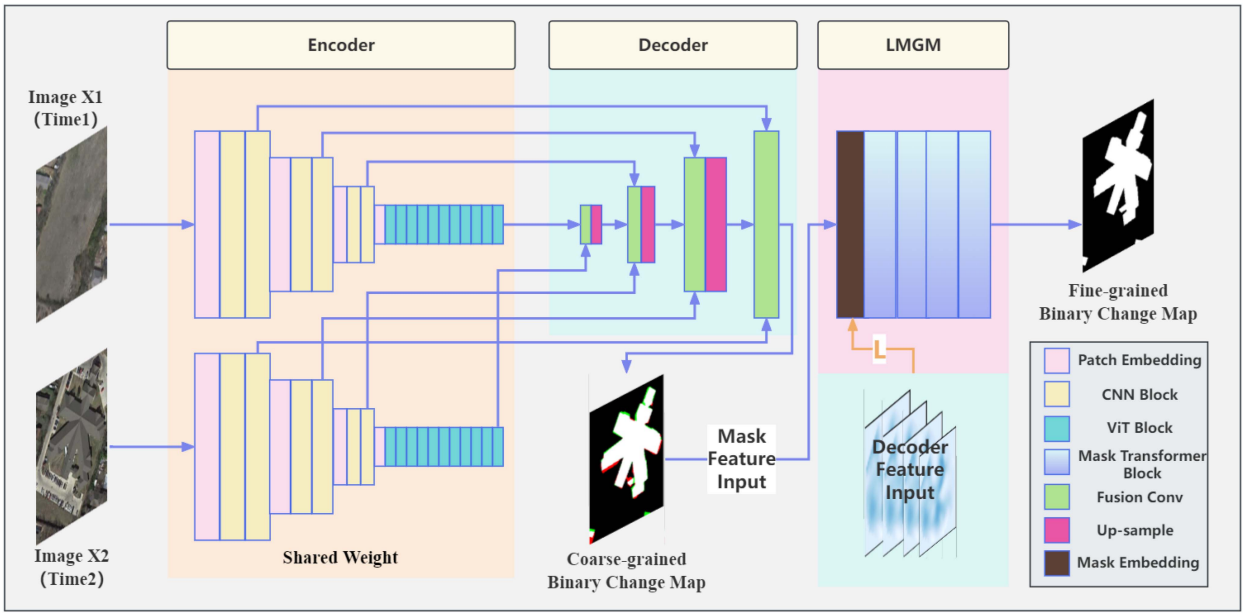

Fig. 3. Overview of the U-Conformer.

A. Overview of the U-Conformer

As shown in Fig. 3, the proposed U-Conformer network, a Siamese network blending CNN and ViT technologies, consists of two primary components: the U-Conformer and the LMGM. The U-Conformer designs a novel architecture incorporating both the Conformer structure, which merges CNN and ViT for enhanced perception of multiscale targets, and a U-shaped Siamese architecture, which synergizes multiscale features from both CNN and ViT to generate robust features for target detection. To improve the model's generalization across diverse architectural targets, the CFMAE is incorporated into the U-Conformer network, exploiting a vast corpus of unlabeled remote sensing images of buildings. This significantly bolsters the feature capture potential of the Conformer. For a more detailed structural explanation, refer to Section III-B. In addition, the LMGM, employing a masked transformer, directs the model's focus toward foreground change information, pooling multiscale features to achieve precise multiscale CD results. Further structural details are provided in Section III-C.

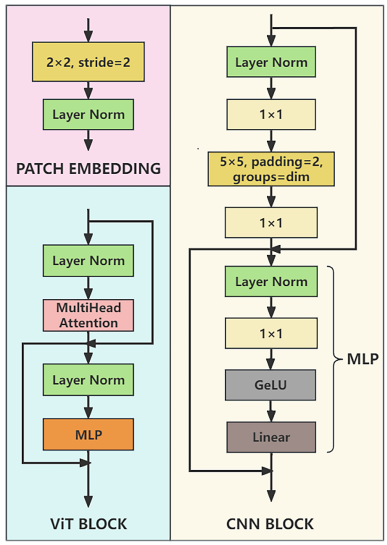

Fig. 4. Detail of the CNN block and the ViT block in the Conformer.

B. U-Conformer

As shown in Fig. 3, to acquire the robust multiscale features necessary for CD, the Conformer is utilized to thoroughly comprehend and encode semantic information at various scales within the given bitemporal remote sensing images, $T_1, T_2 \in \mathbb{R}^{H\times W\times 3}$, which have a spatial resolution of $H \times W$ and a channel number of 3. The specific details of the Conformer are presented in Fig. 4. The system's encoder adopts a Siamese configuration, comprising three stages of CNN blocks followed by a single stage of a Transformer block in each branch. These twin branches operate in a weight-sharing mode, adeptly extracting multiscale features from the input bitemporal images, corresponding to the prechange and postchange scenarios, respectively.

The initial three stages of the convolutional blocks are designed in alignment with the Transformer block's design principles. Instead of employing self-attention mechanisms as found in traditional Transformers, these stages utilize 5×5 depthwise convolutions. This approach is specifically chosen to effectively learn local features from the input images. The functioning within the convolutional block can be articulated as follows:

$$\hat{X}^l = \text{conv}_{1\times1}(\text{conv}_{5\times5}(\text{conv}_{1\times1}(\text{LN}(X^{l-1})))) + X^{l-1} \quad (1)$$

$$X^l = \text{MLP}(\text{LN}(\hat{X}^l)) + \hat{X}^l \quad (2)$$

where $\hat{X}^l$ denotes the output from the convolution module in the lth layer block, and X^l signifies the output produced by the MLP module.

Subsequently, the local features acquired by the CNN stages are channeled into the fourth stage, which employs a Transformer for global semantic modeling. This Transformer stage utilizes self-attention blocks to extract global features. The process of computing self-attention can be described as follows:

$$Q = TW^q, K = TW^k, V = TW^v \tag{3}$$

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where $T \in \mathbb{R}^{M^2 \times d}$ symbolizes the input to the self-attention mechanism, and W^q, W^k, and $W^v \in \mathbb{R}^{d \times d}$ are the weights of three distinct linear projection layers. The matrices $Q, K, V \in \mathbb{R}^{M^2 \times d}$ correspond to the query, key, and value components of the attention model, respectively. Here, M^2 refers to the resolution of the matrices T, Q, K, and V, while d represents the dimension of the individual attention head. The functioning within the Transformer block can be delineated as follows:

$$\hat{X}^l = A(\text{LN}(X^{l-1})) + X^{l-1} \tag{5}$$

$$X^l = \text{MLP}(\text{LN}(\hat{X}^l)) + \hat{X}^l \tag{6}$$

where A symbolizes the self-attention operation. In this formulation, $\hat{X}^l$ denotes the output generated by the transformer module in the lth layer block, while X^l indicates the output produced by the MLP module.

At each transition between stages, patch embedding is employed to downscale the feature map to half its prior spatial resolution. In Stages 1, 2, and 3, the local convolution kernels are designed with comparatively small receptive fields. Conversely, the transformer block in Stage 4 is adept at aggregating and fusing features from more coarsely grained representations, thereby expanding the receptive field to encompass the entire image. This strategic design allows for the effective extraction and integration of both local and global information from the paired images.

After obtaining multiscale feature maps from the bitemporal images, following the concept of the U-shaped network, shallow features are introduced into the decoder via skip connections to enhance the localization accuracy of change information. The decoder is composed of multiple cascaded change feature query blocks, where each change feature query block consists of a 1×1 convolution, a $2\times$ upsampling operation, and a convolutional block. Each convolutional block comprises two repetitions of a 3×3 convolutional layers, a batch normalization layer, and a ReLU layer. Bilinear interpolation is used as the upsampling method. Initially, the decoder compares the high-level semantic features extracted by the Transformer layers in the two encoder branches to decode high-level change features. These are then fused with the upsampled and shallower bitemporal features from the higher level CNN layers. The deep-scale change features guide the decoding of shallow-scale change features, and simultaneously, the shallow-scale change features optimize the recovery of change feature details. Through a series of skip connections, the multiscale bitemporal features from the encoder are repeatedly decoded and fused with the upsampled change features. This ultimately achieves the decoding of high-precision change features we need from the bitemporal building target features, providing a prerequisite for subsequent change information reconstruction.

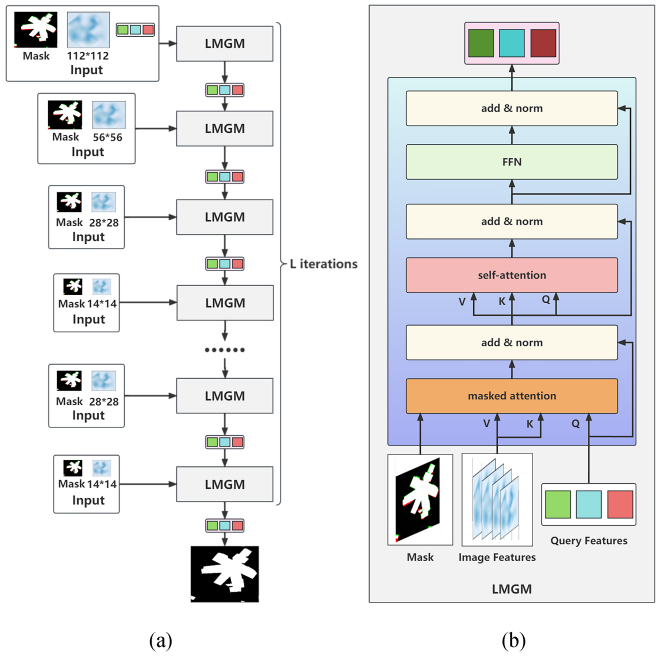

Fig. 5. (a) LMGM process for refining multiscale features. (b) Detailed view of the LMGM.

C. Learnable Mask Guidance Module (LMGM)

The LMGM is an innovative mechanism designed to enhance the performance of the U-Conformer architecture in detecting building changes in remote sensing images, especially in dealing with pseudochange interference and precisely capturing effective change areas, as shown in Fig. 3. This module effectively directs the network's focus to activated regions within multiscale change features, facilitating the refinement of prediction outcomes and resulting in more accurate CD maps. The input to LMGM includes multiscale change features at four different resolutions from the U-Conformer architecture, assisting LMGM in finely processing change information across various scales. In addition, the input incorporates a coarse binary change map output by the U-Conformer architecture as mask information, guiding the algorithm to focus on potential foreground change targets in subsequent learning processes. This design enables the algorithm to effectively locate change areas in early stages, providing a solid foundation for further refinement.

The specific process by which the LMGM method guides the U-Conformer network to focus on foreground change information is illustrated in Fig. 5. Initially, the mask Transformer takes a mask matrix, multiscale change features, and learnable query features as input, as depicted in Fig. 5(b). Here, the mask matrix is generated by masking unactivated areas in the highest

resolution change features output by the U-Conformer. Subsequently, the mask Transformer processes the multiscale change features from the lowest to the highest resolution, under the guidance of the mask matrix, controlling the learnable query features to focus on foreground change information within the multiscale change features, thus achieving refined modification of change information at each scale. Afterwards, the refined features are further optimized through a self-attention mechanism for global modeling. The LMGM layer consists of four mask Transformers, with input feature resolutions of H1 = H/16, H2 = H/8, H3 = H/4, H4 = H/2, and W1 = W/16, W2 = W/8, W3 = W/4, W4 = W/2, where H and W are the original image resolutions. By undergoing L cycles of iterative refinement within the LMGM, the algorithm incrementally refines the change information across different scales, ultimately achieving high-precision CD, as shown in Fig. 5(a). Typically, the standard cross-attention computation is as follows:

$$X_l = \text{softmax}(Q_l K_l^T) V_l + X_{l-1} \tag{7}$$

where l represents the index of the layer, with $X_l \in \mathbb{R}^{N\times C}$ denoting N query features, each of C dimensions, at the lth layer and $Q_l = f_Q(X_{l-1}) \in \mathbb{R}^{N\times C}$. X_0 signifies the initial query features input into the LMGM. The image features transformed by $f_K(\cdot)$ and $f_V(\cdot)$ are represented as $K_l, V_l \in \mathbb{R}^{H_l W_l \times C}$, where H_l and W_l denote the spatial resolution of these image features. The functions f_Q, f_K, and f_V are linear transformations. The masked attention mechanism in the LMGM is adjusted as follows:

$$X_l = \text{softmax}(\hat{M}_{l-1} + Q_l K_l^T) V_l + X_{l-1} \tag{8}$$

where the attention mask $\hat{M}_{l-1}$ at feature location (x, y) is

$$\hat{M}_{l-1}(x,y) = \begin{cases} 0 & \text{if } M_{l-1}(x,y) = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{9}$$

where $M_{l-1} \in \{0,1\}^{N\times H_l W_l}$ represents the binarized (thresholded at 0.5) output of the resized mask prediction of the previous $(l-1)$th layer of the Transformer decoder, and it is resized to match the resolution of K_l. The term M_0 refers to the initial binary mask prediction derived from X_0, which is generated before the query features are input into the LMGM.

By performing change information refinement using only one resolution of change features at a time, the approach avoids interference caused by differences in features at different scales. This enables better multiscale change features fusion and facilitates the effective capture of information at various scales. In addition, by utilizing only one resolution of features for refinement at a time, the method obtains precise binary change maps from bitemporal building change features without significantly increasing the computational complexity of the model. Through LMGM, the U-Conformer network is optimized, further enhancing the accuracy and reliability of detecting building changes in remote sensing images.

D. Loss Function

In the task of detecting building changes in remote sensing images, accurately locating buildings within complex change areas and precisely capturing their change information are crucial for achieving high-precision detection results. Although the advanced U-Conformer architecture and LMGM have to some extent improved the performance of building CD, the inaccuracy of coarse binary change mask information remains a significant challenge they face. This can lead to serious errors in locating change areas, affecting the precision of identifying and locating changed building targets, and thus, causing missed detections or false alarms.

To address this challenge, a mask information joint class-balanced loss function is proposed. The design of this loss function aims to enhance the quality of the coarse binary change mask information through supervised constraints, providing more accurate guidance for CD algorithms. This approach not only improves the algorithm's accuracy in locating change areas but also enhances its ability to recognize changed building targets.

Specifically, the mask information joint class-balanced loss function considers both the accuracy of mask information and the class balance issue in CD during loss calculation. This joint consideration enables the algorithm to focus more on areas with high potential for change while reducing attention to the background or unchanged areas. The loss function ensures the algorithm's robustness in facing class imbalance issues by generating better mask information, especially in scenarios where the number of changed and unchanged samples is highly imbalanced.

In constructing the mask information joint class-balanced loss function, a mask information constraint is first introduced to ensure high-quality mask information. In addition, the binary cross-entropy (BCE) loss function and the Dice coefficient loss function are incorporated, combining the need to address class imbalance with the requirement to guide attention towards the foreground. Ultimately, the mask information joint class-balanced loss function combines the mask information constraint, the BCE loss and the Dice coefficient loss, setting appropriate weights for these loss components to further optimize the model's detection performance. The expression for the mask information joint class-balanced loss function is as follows:

$$L = L_{\text{mask}} + \lambda_1 L_{\text{bce}} + \lambda_2 L_{\text{dice}} \tag{10}$$

where L_{mask} represents the mask information constraint, which entails using BCE loss on changing features before performing feature masking operations, to ensure the accuracy of mask information. L_{bce} signifies the BCE loss, while L_{dice} represents the dice coefficient loss. The terms λ_1 and λ_2 are used to denote the balancing weight between these three types of losses. In addition, the mask information constraint is delineated as follows:

$$L_{\text{mask}} = L_{\text{bce}}(F_{\text{mask}}, F_{\text{GT}}) \tag{11}$$

where F_{mask} and F_{GT} represent the coarse binary change map generated from change features before the feature masking operation and the change ground-truth image, respectively. The

BCE loss is computed as follows:

$$L_{\text{bce}} = -(y\log(p(x)) + (1-y)\log(1-p(x))) \tag{12}$$

where $p(x)$ and y represent the predicted image and the ground-truth image. The dice coefficient is defined as the similarity between two contour regions, defined as

$$L_{\text{dice}} = 1 - \frac{2y_i t_i}{y_i + t_i} \tag{13}$$

where y_i denotes the predicted probability of a pixel being classified as part of the changed category, while t_i represents the corresponding ground-truth label for that pixel.

Ultimately, by constructing and utilizing the mask information joint class-balanced loss function, the BCD network demonstrates a more powerful capability in capturing change information, showcasing exceptional detection performance for building change targets across various scales. This approach combines mask information guidance with class-balancing methods, meeting the needs of both U-Conformer and LMGM, and represents a significant technical advancement in the field of BCD in remote sensing images.

IV. Experiments and Analysis

To investigate the effectiveness of the proposed U-Conformer, three publicly available and challenging BCD datasets, namely LEVIR-CD, WHU-CD, and GZ-CD, were utilized in the experiments. In this section, the following aspects will be presented. First, a detailed description of these datasets will be provided. Second, the evaluation metrics used and several SOTA CD methods will be showcased. Third, implementation details of these methods will be provided. Subsequently, a series of ablation studies will be deployed and discussed. Finally, a comparison and analysis of experimental results based on the three BCD datasets will be presented.

A. Dataset Descriptions

In the experiments, three BCD datasets were used to evaluate the performance of the proposed U-Conformer. The detailed information about these datasets is as follows.

1) *LEVIR-CD Dataset (Google Earth Image) [21]:* The LEVIR-CD dataset is a publicly available dataset specifically designed for BCD. It includes a variety of building types such as villas, high-rise apartments, small garages, and large warehouses. Thus, the diverse forms and arrangements of buildings in LEVIR-CD images pose significant challenges for BCD algorithms in accurately recognizing different types of structures.
2) *WHU-CD Dataset (Aerial Image) [37]:* The WHU-CD dataset is another publicly available dataset specifically tailored for BCD. It captures various building changes, including demolished and reconstructed buildings, as well as diverse ground objects such as grass, trees, roads, bridges, vehicles, and parking lots. The buildings exhibit significant differences in scale, shape, and color, and can easily be confused with other ground objects, which demands that BCD algorithms have strong capabilities to distinguish varying building targets.
3) *GZ-CD Dataset (Google Earth Image) [38]:* The GZ-CD dataset is a publicly available BCD dataset created by Peng et al.[38]. It consists of 19 pairs of seasonal change images covering the suburban area of Guangzhou, China, from 2006 to 2019. The buildings exhibit significant shape and size variations, ranging from large industrial and residential buildings to small mobile residences. In addition, the GZ-CD dataset contains a large number of pseudochange interferences caused by lighting and weather conditions, with particularly large displacements caused by the perspective projection of high-rise buildings. Finally, the clarity of the GZ-CD dataset is lower compared to the first two datasets, making it exceptionally challenging.

B. Evaluation Metrics and Comparative Methods

1) *Evaluation Metrics:* This article employs six widely recognized evaluation metrics to quantitatively assess the U-Conformer model's performance from various angles. These metrics include overall accuracy (OA, %), intersection over union (IoU, %), precision (PRE, %), recall (REC, %), F1-Score (F1, %), and algorithm execution time (Time, seconds/pair). They provide a multidimensional view of the model's effectiveness, where the algorithm execution time represents the time required to predict a pair of remote sensing images. True positive (TP) in this study is defined as the number of pixels accurately identified as buildings that have undergone changes. true negative (TN) refers to the pixels correctly classified as either unchanged buildings or nonbuilding entities. False positive (FP) is the count of pixels mistakenly identified as changed buildings, and false negative (FN) is the number of pixels that were not detected as changed buildings but should have been. Utilizing these definitions, the aforementioned metrics are calculated as follows, providing a comprehensive evaluation of the U-Conformer model:

$$\text{OA} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{14}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{15}$$

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

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

$$F1 = \frac{2\times\text{PRE}\times\text{REC}}{\text{PRE}+\text{REC}}. \tag{18}$$

2) *Comparative Methods:* To validate the effectiveness of the proposed U-Conformer method, a comprehensive and thorough comparison was conducted in the experiments with ten related and SOTA methods. These methods are categorized into three groups based on their underlying technology: CNN-based algorithms, Transformer-based

algorithms, and those that integrate both CNN and Transformer architectures.

a) *CNN-based algorithms:* These included FC-Siam-Conc [20], FC-Siam-Diff [20], DASNet [21], STANet [22], BSFNet [58], SNUNet [59], and DMINet [25]. For these algorithms, except for FC-Siam-Conc and FC-Siam-Diff, we consistently employed ResNet50 as the encoder, incorporating officially released pretrained parameters to maximize their performance. This strategy ensures a robust and equitable comparison with our U-Conformer approach.

b) *Transformer-based algorithm:* We included ChangeFormer [33] in our evaluation and utilized the pretrained parameters supplied by its original developers to maintain consistency in performance evaluation.

c) *CNN-Transformer fusion algorithms:* This category comprised BITNet [36] and MSCANet [60]. We also leveraged the pretrained parameters as provided by the original authors, ensuring that these models perform at their best for a fair comparison with our U-Conformer.

This diverse selection of algorithms provides a comprehensive benchmark, allowing us to thoroughly evaluate the U-Conformer's performance against a spectrum of SOTA approaches in the field.

C. Implementation Details

In our experimental setup, the model was developed using PyTorch and trained on a Tesla V100 PCIe 32-GB GPU. The training data involved segmenting the images into smaller patches, each measuring 224 $\times$ 224 pixels. For the LEVIR-CD dataset, this approach yielded 11 125 training pairs, 1 600 pairs for validation, and 3 200 pairs for testing. In the case of the WHU-CD dataset, we excluded images lacking change targets, resulting in 1 582 training pairs, 226 for validation, and 452 for testing. A similar criterion was applied to the GZ-CD dataset, leading to 1 203 training pairs, 144 validation pairs, and 144 test pairs. To prevent overfitting, random data augmentation techniques were applied, including vertical and horizontal flips, along with random rotations. The model was trained using the AdamW optimizer, set with a weight decay of 0.0001. The initial learning rate was 0.0001, which was linearly reduced to zero over 150 training epochs for LEVIR-CD and WHU-CD datasets, and over 80 epochs for the GZ-CD dataset. A batch size of 12 was employed throughout the training process.

D. Experimental Comparison and Analysis

To evaluate the superiority of the proposed U-Conformer, we compared it with ten classic or SOTA methods, including FC-Siam-Diff, FC-Siam-Conc, DASNet, STANet, BITNet, BSFNet, SNUNet, ChangeFormer, MSCANet, and DMINet, on the LEVIR-CD, WHU-CD, and GZ-CD datasets. These methods encompass various fusion techniques, such as feature difference and feature concatenation, commonly used in CD networks. In addition, they extensively explore the feasibility of utilizing attention modules and contrastive learning to assist in feature learning. The comparative methods also include pure CNN and pure Transformer networks to investigate the advantages and limitations of each. Moreover, hybrid networks that combine both CNN and Transformer architectures are also included in the comparison to investigate their potential. In this section, we conducted comprehensive experimental comparisons and analyses of these methods.

1) Results on the LEVIR-CD Dataset: In the LEVIR-CD dataset, the presence of numerous buildings with unique forms necessitates high capabilities for detail description. As a result, networks like FC-Siam-Conc and DMINet, which use convolutional architectures, achieve better prediction results than ChangeFormer, a Transformer-based network, due to the convolutional architecture's stronger local feature capturing ability. However, due to the limited receptive field of convolutions, CNNs are less effective at utilizing contextual information. This limitation results in poorer performance in discriminating changed building targets compared to our proposed U-Conformer, leading to more false alarms and missed detections. A numerical comparison of the different methodologies is detailed in Table I. Comparing the quantitative results with different methods, the proposed U-Conformer achieves the highest accuracy in terms of OA, IoU, and F1 (99.1%, 84.3%, and 91.5%). Significant improvements are observed over other methods. For instance, compared to DASNet, which had the poorest performance, U-Conformer shows improvements of approximately 0.8%, 11.6%, and 7.3% in the three comprehensive evaluation metrics (OA, IoU, and F1), respectively. Compared to DMINet, which had the best performance, improvements of 0.1%, 2.2%, and 1.3% are achieved in OA, IoU, and F1, respectively. In terms of algorithm execution time, methods employing multilayer Transformer structures (such as ChangeFormer and U-Conformer) tend to require more time. However, our approach, by integrating with CNNs, achieves faster operational efficiency and higher detection accuracy compared to networks with purely Transformer architectures.

In addition to quantitative metrics, visual results also substantiate the advantages of our method. The building change maps (BCMs) obtained through various methods on the LEVIR-CD dataset are shown in Fig. 6, depicting representative examples. By observing the examples, it is evident that the proposed U-Conformer provides clearer BCMs for buildings with different shapes and scales. For example, in the last group of scenes, where there are numerous small building targets closely arranged, U-Conformer can accurately detect almost all targets and precisely delineate building edges. In contrast, other methods miss two or more building targets, with failing to accurately outline building edges, resulting in numerous false alarms. In addition, in the first, third, and fifth groups of scenes featuring large or irregularly changing buildings, only our proposed U-Conformer achieves relatively precise BCMs. In contrast, convolutional networks such as DASNet and STANet suffer from severe false alarms and missed detections, almost failing to accurately locate and recognize changing building targets. Meanwhile, Transformer networks like ChangeFormer exhibit significant missed detections in the center of the building, presenting pixel voids and failing to fully describe the changing architectural area details. Finally, hybrid architecture networks such as BITNet and MSCANet, due

TABLE I
QUANTITATIVE COMPARISON RESULTS (IN%) OF VARIOUS METHODS APPLIED ON THE LEVIR-CD DATASET

Methods	Magazine	Publication Date	Backbone	OA	IoU	REC	PRE	F1	Time
FC-Siam-Diff [20]	ICIP	2018	/	98.9	80.2	86.0	92.2	89.0	0.7
FC-Siam-Conc [20]	ICIP	2018	/	98.9	80.8	87.6	91.3	89.4	1.1
DASNet [21]	TGRS	2020	Resnet50	98.3	72.7	88.7	80.2	84.2	0.9
STANet [22]	Remote Sensing	2020	Resnet50	98.6	75.5	80.8	91.9	86.0	0.5
BITNet [36]	TGRS	2021	Resnet50	98.7	77.9	85.9	89.3	87.6	0.2
BSFNet [58]	GRSL	2021	Resnet50	98.8	79.5	89.7	87.6	88.6	0.3
SNUNet [59]	GRSL	2021	Resnet50	99.0	81.4	87.8	91.8	89.7	1.8
ChangeFormer [33]	IGARSS	2022	Resnet50	98.9	80.6	87.2	91.5	89.3	4.0
MSCANet [60]	J-STARS	2022	Resnet50	98.9	81.0	87.6	91.5	89.5	0.2
DMINet [25]	TGRS	2023	Resnet50	99.0	82.1	89.5	90.9	90.2	0.2
U-Conformer (our)	/	/	Conformer	**99.1**	**84.3**	**90.7**	**92.2**	**91.5**	3.2

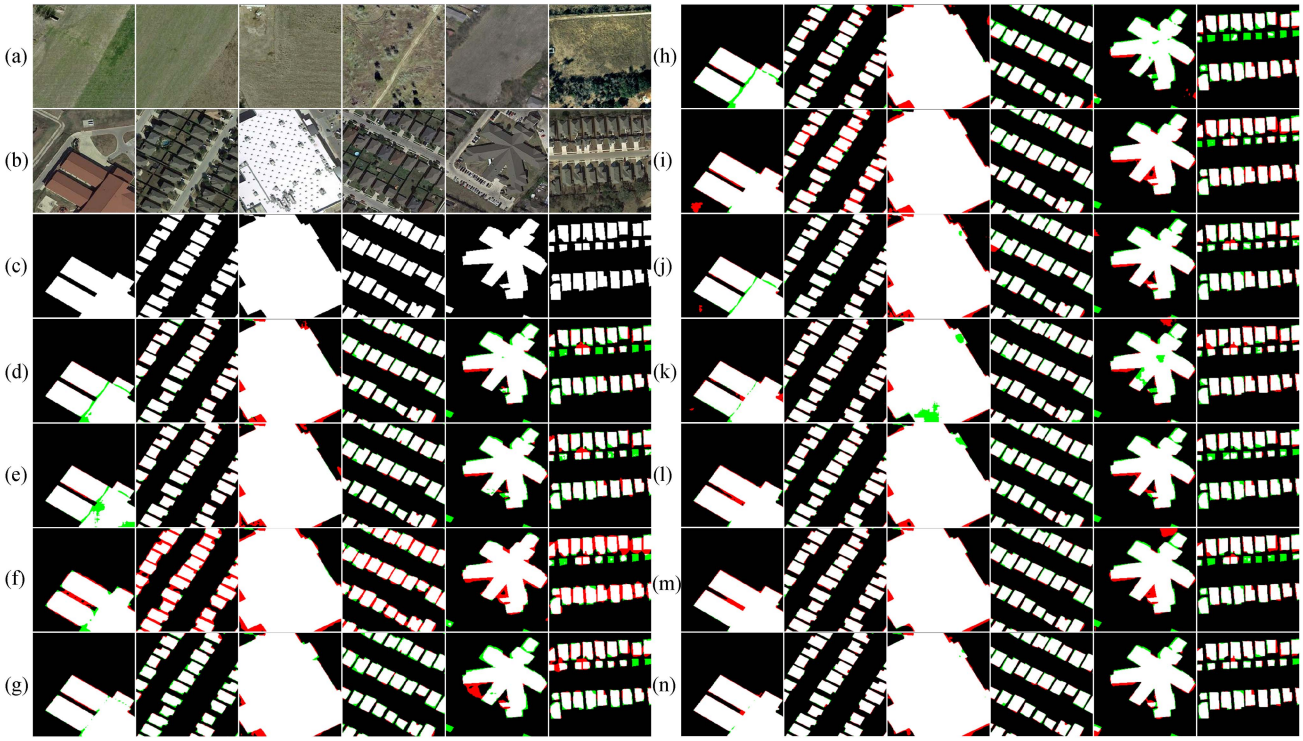

Fig. 6. Representative examples of visualization results of various methods on the LEVIR-CD dataset. (a) T1-time image. (b) T2-time image. (c) GT map. (d) FC-Siam-Dif. (e) FC-Siam-Conc. (f) DASNet. (g) STANet. (h) BITNet. (i) BSFNet. (j) SNUNet. (k) ChangeFormer. (l) MSCANet. (m) DMINet. (n) Proposed U-Conformer. (Notations: green, red, white, and black denote missed detection pixels, false detection pixels, correct detection changed pixels, and correct detection unchanged pixels, respectively.)

to their insufficient handling of convolutional local features and ViT global features, result in suboptimal detection outcomes. Notably, for the small target in the lower left corner of scene five, only our proposed U-Conformer successfully detects it, whereas all other methods fail. In the dense target predictions of the second and fourth group of scenes, U-Conformer remains stable and accurate in detecting all targets and describes more precise target boundaries compared to other methods. Overall, our proposed U-Conformer achieves the best performance on the LEVIR-CD dataset.

2) Results on the WHU-CD Dataset: In the WHU-CD dataset, the presence of numerous nonbuilding objects resembling buildings significantly interferes with the network's ability to recognize building targets. An effective BCD network requires an enhanced capability to extract building features. Traditional convolutional networks such as FC-Siam-Diff, FC-Siam-Conc, and DASNet have been unsatisfactory in feature extraction, significantly underperforming compared to Transformer networks like ChangeFormer. However, convolutional networks can achieve substantial improvements, and even surpass Transformer network performance, by incorporating appropriate attention mechanisms, as seen with networks like BSFNet. Therefore, when the U-Conformer appropriately merges convolutional and Transformer structures, it naturally achieves the best results. A numerical comparison of the different methodologies is detailed in Table II. Comparing the quantitative results with

TABLE II
QUANTITATIVE COMPARISON RESULTS (IN%) OF VARIOUS METHODS APPLIED ON THE WHU-CD DATASET

Methods	Magazine	Publication Date	Backbone	OA	IoU	REC	PRE	F1	Time
FC-Siam-Diff [20]	ICIP	2018	/	86.5	54.7	84.0	61.0	70.7	0.5
FC-Siam-Conc [20]	ICIP	2018	/	82.8	49.6	87.5	53.4	66.3	0.5
DASNet [21]	TGRS	2020	Resnet50	81.8	48.5	88.6	51.8	65.4	0.2
STANet [22]	Remote Sensing	2020	Resnet50	90.0	60.8	79.9	71.8	75.6	0.3
BITNet [36]	TGRS	2021	Resnet50	87.9	58.7	88.8	63.4	74.0	0.2
BSFNet [58]	GRSL	2021	Resnet50	95.8	81.3	**94.5**	85.4	89.7	0.2
SNUNet [59]	GRSL	2021	Resnet50	91.7	67.6	89.6	73.3	80.7	1.7
ChangeFormer [33]	IGARSS	2022	Resnet50	95.4	78.5	87.0	88.9	87.9	2.5
MSCANet [60]	J-STARS	2022	Resnet50	95.5	79.3	88.5	88.5	88.5	0.5
DMINet [25]	TGRS	2023	Resnet50	90.0	63.6	89.9	68.5	77.7	0.2
U-Conformer (our)	/	/	Conformer	**97.9**	**89.7**	93.9	**95.3**	**94.6**	0.8

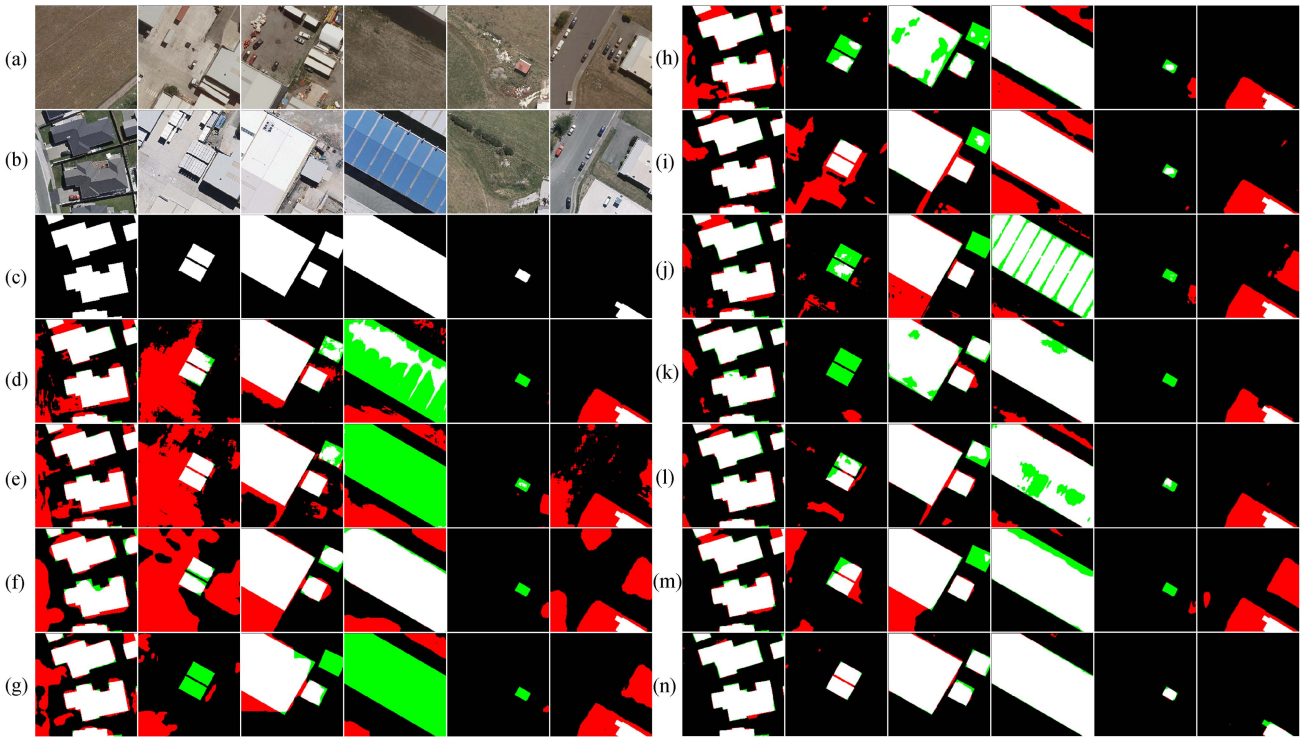

Fig. 7. Representative examples of visualization results of various methods on the WHU-CD dataset. (a) T1-time image. (b) T2-time image. (c) GT map. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) DASNet. (g) STANet. (h) BITNet. (i) BSFNet. (j) SNUNet. (k) ChangeFormer. (l) MSCANet. (m)DMINet. (n) Proposed U-Conformer. (Notations: green, red, white, and black denote missed detection pixels, false detection pixels, correct detection changed pixels, and correct detection unchanged pixels, respectively.)

different methods, the proposed U-Conformer achieves the highest accuracy in terms of OA, IoU, and F1 (97.9%, 89.7%, and 94.6%). Significant improvements are observed over other methods. For instance, compared to DASNet, which had the poorest performance, U-Conformer shows improvements of approximately 16.1%, 41.2%, and 29.2% in the three comprehensive evaluation metrics (OA, IoU, and F1), respectively. Compared to BSFNet, which had the best performance, improvements of 2.1%, 8.4%, and 4.9% are achieved in OA, IoU, and F1, respectively. When processing relatively simple remote sensing images, the U-Conformer significantly reduces execution time, even outperforming complex pure CNNs (such as SNUNet) in terms of speed.

In addition to quantitative metrics, visual results also substantiate the advantages of our method. BCMs obtained through various methods on the WHU-CD dataset are shown in Fig. 7, displaying representative examples. Despite the relatively smaller sample size compared to the LEVIR-CD dataset, the proposed U-Conformer still provides BCMs with significantly higher accuracy than other methods. U-Conformer demonstrates superior identification results for various building targets, regardless of large-scale factories or small-scale farmhouses, urban garden villas, or industrial cement floor buildings. For example, in the first four scenes of Fig. 7, U-Conformer accurately identifies targets with different shapes and scales, providing precise descriptions of building edges. Meanwhile, other methods produce

TABLE III
QUANTITATIVE COMPARISON RESULTS (IN%) OF VARIOUS METHODS APPLIED ON THE GZ-CD DATASET

Methods	Magazine	Publication Date	Backbone	OA	IoU	REC	PRE	F1	Time
FC-Siam-Diff [20]	ICIP	2018	/	90.4	66.6	79.8	80.0	79.9	0.7
FC-Siam-Conc [20]	ICIP	2018	/	90.7	67.0	79.0	81.5	80.2	0.7
DASNet [21]	TGRS	2020	Resnet50	87.7	56.7	67.3	78.3	72.4	0.9
STANet [22]	Remote Sensing	2020	Resnet50	86.4	47.3	51.0	86.9	64.3	0.3
BITNet [36]	TGRS	2021	Resnet50	90.9	67.0	77.3	83.5	80.3	0.4
BSFNet [58]	GRSL	2021	Resnet50	91.9	69.5	76.8	**88.0**	82.0	0.2
SNUNet [59]	GRSL	2021	Resnet50	91.2	68.6	79.8	83.0	81.4	1.6
ChangeFormer [33]	IGARSS	2022	Resnet50	89.1	61.8	73.6	79.3	76.4	1.0
MSCANet [60]	J-STARS	2022	Resnet50	92.3	72.8	85.9	82.6	84.2	0.1
DMINet [25]	TGRS	2023	Resnet50	90.0	64.9	77.5	80.0	78.7	0.2
U-Conformer (our)	/	/	Conformer	**93.7**	**76.5**	**86.2**	87.2	**86.7**	0.7

varying degrees of false alarms and missing detections, especially for complex scenes with building targets, resulting in a large number of errors. In the prediction for the farmhouses in scene five, our U-Conformer accurately detects the targets and precisely describes the building boundaries, showcasing good performance in small target detection. Among other methods, only CNNs like BSFNet and SFNet, which introduced appropriate attention mechanisms, and hybrid architecture networks like BITNet and MSCANet, detected the house in scene five, but none were able to fully delineate its boundaries. In scene six, where changed building targets appear on newly laid cement floors, significant interference occurs, leading to extensive false alarms in other methods. However, our U-Conformer accurately detects the building targets without producing false alarms, demonstrating its more effective discriminative ability for building targets. Overall, our proposed U-Conformer achieves the best performance on the WHU-CD dataset.

3) Results on the GZ-CD Dataset: In the GZ-CD dataset, due to changes in imaging angles and seasonal lighting, there are numerous instances of pseudochange interference. Pure convolutional or pure Transformer networks such as FC-Siam-Diff, DASNet, STANet, and ChangeFormer, due to their lack of local-global information integration capabilities, underperform compared to BITNet, which simply fuses shallow convolutional layers with ViT structures. MSCANet, which better combines convolutional and ViT structures, achieves excellent results, demonstrating the powerful local-global representational capabilities of hybrid structures in handling pseudochange interference with strong robustness. At the same time, networks with better attention mechanisms, such as SNUNet and BSFNet, also gain some capacity to integrate local-global information, thus effectively resisting pseudochange interference and achieving satisfactory prediction results. Our proposed U-Conformer, however, continues to excel with its powerful capability to merge local and global information, resulting in the best performance. A numerical comparison of the different methodologies is detailed in Table III. Compared to different methods, the proposed U-Conformer achieves the highest accuracy in terms of OA, IoU, and F1 (93.7%, 76.5%, and 86.7%), showcasing significant improvements. For example, compared to STANet, which had the poorest performance, U-Conformer shows improvements of approximately 7.3%, 29.2%, and 22.4% in the three comprehensive evaluation metrics (OA, IoU, and F1), respectively. Compared to MSCANet, which had the best performance, improvements of 1.4%, 3.7%, and 2.5% are achieved in OA, IoU, and F1, respectively. When handling simple remote sensing images, the U-Conformer still operates with less execution time compared to complex pure CNNs (such as DASNet and SNUNet) and pure Transformer networks (like ChangeFormer).

In addition to quantitative metrics, visual results also substantiate the advantages of our method. BCMs obtained through various methods on the GZ-CD dataset are illustrated with representative examples in Fig. 8. Despite the significantly smaller sample size compared to the other two datasets and the challenges posed by the 19 groups of remote sensing images captured under different conditions, our proposed U-Conformer still excels in completing the detection task. In various scenes, compared to other methods, U-Conformer exhibits lower rates of false alarms and missing detections, resulting in better recognition results for various complex building targets. For instance, in scene one, other methods suffer from a significant number of false alarms and missed detections, particularly in detecting the elongated building in the upper left corner, with no method accurately describing it. In contrast, our U-Conformer effectively detects four targets of varying scales in scene one. In scenes two, three, and four, due to spatial displacement caused by the remote sensing imaging angle and the 3-D shape of buildings, as well as severe shadow interference from lighting, other methods all exhibit numerous false alarms and missed detections. Specifically, in scene three, only STANet, BSFNet, and our U-Conformer successfully avoid false alarms caused by displacement interference due to the different imaging angles of the building in the upper left corner. However, the substantial missed detections by STANet and the poorer building boundary description by BSFNet prove the superior detection performance of our U-Conformer. In scene five, U-Conformer surpasses other methods with lower rates of false alarms and missed detections for three building targets. In scene six, compared to other methods, U-Conformer significantly reduces false alarms, showcasing its outstanding performance in complex scenarios. Overall, our proposed U-Conformer achieves the best performance on the GZ-CD dataset.

In summary, compared to the selected ten methods, our proposed U-Conformer achieved superior results on the LEVIR-CD

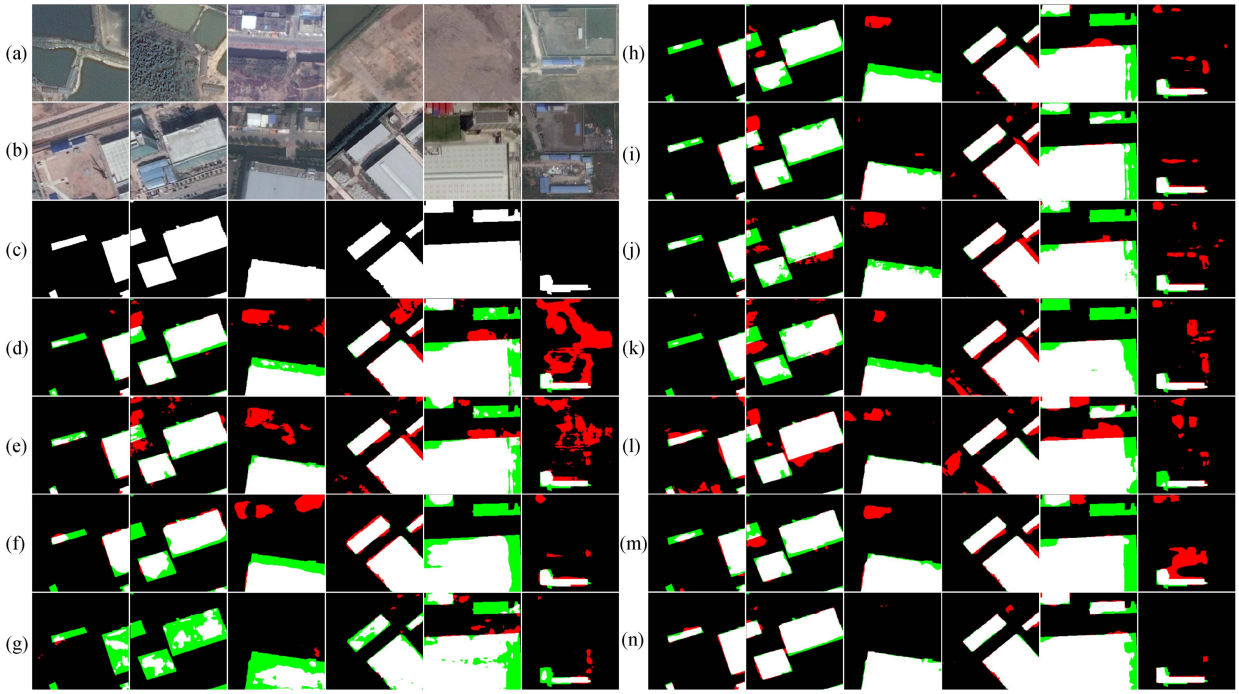

Fig. 8. Representative examples of visualization results of various methods on the GZ-CD dataset. (a) T1-time image. (b) T2-time image. (c) GT map. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) DASNet. (g) STANet. (h) BITNet. (i) BSFNet. (j) SNUNet. (k) ChangeFormer. (l) MSCANet. (m)DMINet. (n) Proposed U-Conformer. (Notations: green, red, white, and black denote missed detection pixels, false detection pixels, correct detection changed pixels, and correct detection unchanged pixels, respectively.)

dataset, WHU-CD dataset, and GZ-CD dataset. Based on the experimental results, the following conclusions can be drawn.

1) U-Conformer has been extensively validated to exhibit robust discriminative capabilities for detecting building targets of various scales and shapes, yielding more accurate details of building objects. The Conformer, along with the LMGM, effectively fuse long-range semantic information with local details, resulting in precise modeling of buildings with diverse scales and shapes. This integration enables the network to capture both global context and fine-grained features, contributing to its exceptional performance in BCD tasks. U-Conformer has demonstrated superior performance over other methods through extensive and thorough experimentation, illustrating its exceptional capability in managing complex and demanding scenarios with notable precision and effectiveness.
2) U-Conformer exhibits exceptional resilience against false alarms and omissions, surpassing the performance of other methods significantly. Its robustness is particularly evident in its ability to handle diverse environmental factors, including lighting variations, seasonal changes, and complex scenarios that might induce pseudochanges. The LMGM effectively focuses the network's attention on foreground objects at different scales, continually learning to discern various scales of building targets. As a result, U-Conformer's resistance to different types of interference is greatly enhanced, enabling it to accurately detect real building changes amidst challenging and dynamic conditions.
3) U-Conformer continues to excel on both the GZ-CD and WHU-CD datasets, even with limited sample sizes. It effectively identifies and accurately describes various building targets, showcasing its proficiency in few-shot learning tasks and reinforcing its generalization capability. The CFMAE proves to be highly advantageous as it utilizes a substantial number of unlabeled remote sensing images for feature learning. This approach significantly contributes to the network's enhanced generalization performance when compared to other pretraining methods.

Overall, the results indicate that U-Conformer is a powerful and robust method for BCD, surpassing existing SOTA approaches across various datasets.

E. Ablation Study on WHU-CD Dataset

To effectively tackle the challenges in BCD tasks, we have designed corresponding modules targeting different difficulties. First, to address the scale and shape variations of building targets, we devised the U-Conformer to capture more robust change features, especially for large buildings and dense small clusters of buildings. In addition, to handle the significant intraclass differences of buildings and the interference from complex environments, we designed the LMGM to suppress pseudochange interference and focus the model on foreground targets. Furthermore, we proposed the CFMAE to enhance the model's generalization, enabling it to adapt to various target scenarios. In order to validate the effectiveness of the main network backbone and components, we conducted three ablation studies on the

TABLE IV
QUANTITATIVE EVALUATION ACCURACY (IN%) FOR ABLATION STUDY OF EACH SUBMODULE ON THE WHU-CD DATASET

Submodel Name	Conformer (M1)	U-Conformer (M2)	LMGM without L_{mask} (M3)	LMGM with L_{mask} (M4)	CFMAE (M5)	OA	IoU	F1	Time
M1	✔					95.4	78.9	88.2	1.1
M1+M2	✔	✔				96.1	82.3	90.3	0.8
M1+M2+M3	✔	✔	✔			96.3	81.6	89.9	0.8
M1+M2+M4	✔	✔		✔		97.3	87.0	93.1	0.8
M1+M2+M4+M5	✔	✔		✔	✔	**97.9**	**89.7**	**94.6**	0.8

WHU-CD dataset, as described below. In addition, this section utilized four comprehensive metrics (OA, IoU, F1, and Time) to quantitatively evaluate the results of the ablation experiments.

1) Ablation Study for Each Submodule: To investigate the impact of each submodule, the proposed U-Conformer architecture is decomposed into four submodules: the Conformer, the U-Conformer, the LMGM without L_{mask}, the LMGM with L_{mask}, and the CFMAE. The Conformer indicates that only the Convolution-Transformer Siamese network is used as the building feature extractor, instead of the proposed Convolution-Transformer U-Net. In this ablation study, each submodule is gradually combined, and a comparison is performed for the BCD task. As shown in Table IV, the results for different combinations of submodules are evaluated on the WHU-CD dataset. When the U-shaped architecture are designed into the Conformer for BCD, improvements are observed in OA, IoU, and F1 by 0.7%, 3.4%, and 2.1%, respectively. Subsequently, the effectiveness of the mask information constraint is validated. After proposing the LMGM with loss, the model shows improvements of 1.2%, 4.7%, and 2.8% in terms of OA, IoU, and F1, respectively, compared to U-Conformer alone. Meanwhile, we compared the scenario of the LMGM without Loss, where in this case, the model experiences a decrease in IoU and F1 relative to U-Conformer alone, highlighting the instability of the guidance effect. This indicates that ensuring the effectiveness and high quality of mask information is essential to fully unleash the potential of LMGM. Moreover, using the CFMAE to accelerate network convergence and improve network generalization achieves further enhancements in OA, IoU, and F1 by 0.6%, 2.7%, and 1.5%, respectively. In terms of algorithm execution time, the stacking of various modules has hardly imposed any burden on the algorithm, with no significant change in execution time (note: The reduction in execution time from M1 to M2 is primarily due to the removal of numerous convolution operations in the M1 decoder). Finally, combining all submodules into the proposed U-Conformer achieves the best accuracy on the WHU-CD dataset (97.9%, 89.7%, and 94.6%). These quantitative results indicate that the combination of the U-shaped architecture, the LMGM complete with L_{mask}, and the CFMAE can effectively improve the model's performance. The visual results of each submodule in this ablation study also provide representative examples that align with the quantitative findings, as shown in Fig. 9. Notably, scenario 3 illustrates how incorrect mask information can lead to the loss of targets, which matches our experimental quantitative outcomes. Moreover, the LMGM's effect in suppressing pseudochanges is significant, greatly reducing the false positive and missrates in prediction outcomes.

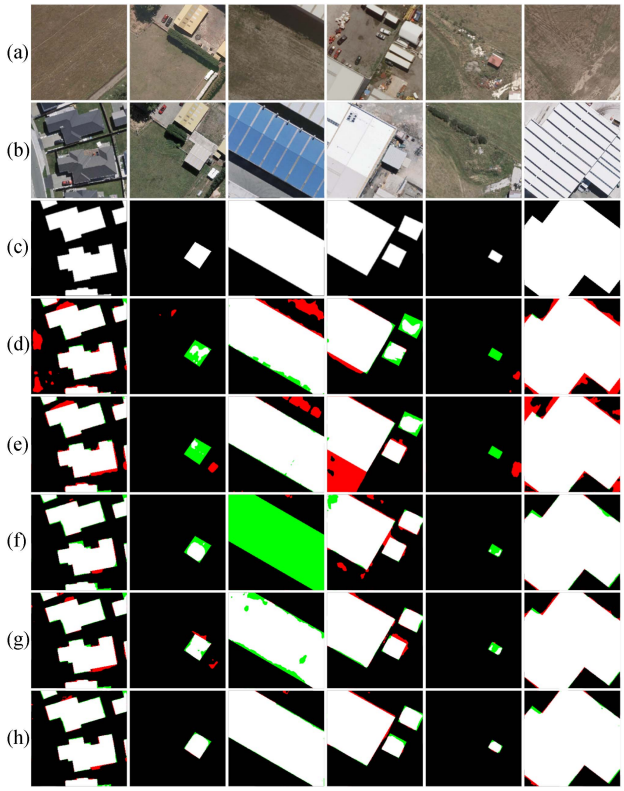

Fig. 9. Representative examples of visualization results for each submodule on the WHU-CD dataset. (a) T1-time image. (b) T2-time image. (c) GT map. (d) Conformer. (e) U-Conformer. (f) U-Conformer + LMGM without L_{mask}. (g) U-Conformer + LMGM with L_{mask}. (h) U-Conformer + LMGM with L_{mask} + CFMAE. (Notations: green, red, white, and black denote missed detection pixels, false detection pixels, correct detection changed pixels, and correct detection unchanged pixels, respectively.)

In addition, the other submodules all contribute substantial gains to the model's predictions.

2) Ablation Study for Mask Matrix Generation in the LMGM: To investigate the influence of various scale change features in U-Conformer on the accuracy of the mask information matrix, we conducted four experiments, namely: using only the first-layer CNN features, using a fusion of first-layer and second-layer CNN features, using fusion of all CNN layers' features, and using fusion of all CNN+Transformer layers' features. In this ablation study, features at different resolutions were progressively integrated, and comparisons were made for the BCD task. The results of this ablation study are obtained on the WHU-CD dataset, and the quantitative evaluation for the entire dataset is shown in Table V. The fusion method of using fusion of

TABLE V
QUANTITATIVE EVALUATION ACCURACY (IN%) FOR ABLATION STUDY OF MASK MATRIX GENERATION STRATEGIES ON THE WHU-CD DATASET

Mask Matrix Generation Strategy	Stage1 in CNN (L1)	Stage2 in CNN (L2)	Stage3 in CNN (L3)	Transformer Stage (L4)	OA	IoU	F1
L1	✔				95.6	78.5	87.9
L1+L2	✔	✔			97.4	87.6	93.4
L1+L2+L3	✔	✔	✔		96.5	82.6	90.5
L1+L2+L3+L4	✔	✔	✔	✔	**97.9**	**89.7**	**94.6**

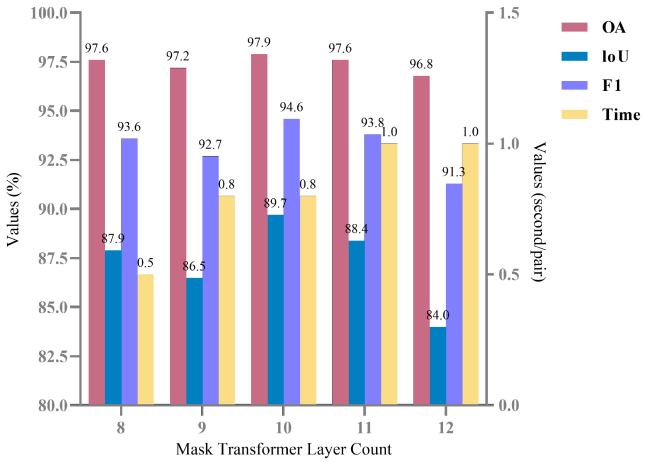

Fig. 10. Quantitative evaluation accuracy (IN% and second/pair) for ablation study of LMGM mask Transformer layer count on the WHU-CD dataset.

all CNN+Transformer layers' features achieves more accurate BCD results.

3) Ablation Study for the Iteration Count of the LMGM: The LMGM learns the target change situation at the current scale by processing the input change features of different scales together with the mask information, and it generates new, optimized change features. These features are then used to update the mask information. After continuous optimization with change features of different scales, the accuracy of the mask information for different scales of change improves progressively. Therefore, in addition to the accuracy of the mask information in the LMGM being related to the number of iterations of the LMGM, the quality of the change features also improves as multiscale optimization progresses. To investigate the impact of the iteration count of the LMGM on the detection outcomes for BCD, we configured varying iteration counts and multiscale feature input layers for the LMGM, tailored to the building scale scenarios in the dataset. This approach was employed to identify the most suitable configuration results. The results of this ablation study are obtained on the WHU-CD dataset, and the quantitative evaluation for the entire dataset is shown in Fig. 10. As the number of iterations increases, the execution time of the algorithm also gradually increases. However, it has been observed that the algorithm achieved its optimal detection results after conducting ten LMGM iterations. This not only indicates that high-precision BCD can be achieved by repetitively leveraging change features of various scales for change information reconstruction, but it also demonstrates that excessively stacking Transformer layers does not always lead to improvements in the algorithm. Considering that prolonged and repeated updates of mask information can lead to overfitting, it is essential to select configuration conditions that are most suitable for the current task.

In conclusion, the ablation studies have explored the impact of different components and configurations in U-Conformer on BCD performance. The results confirm the effectiveness of the U-conformer, LMGM with L_{mask}, CFMAE, and ten iterations of the LMGM in achieving accurate BCD results on the BCD dataset.

V. CONCLUSION

This article designs the U-Conformer, a cutting-edge architectural design tailored specifically for BCD. U-Conformer is founded on an encoder–decoder structure, utilizing the Conformer to extract architectural features. These features are subsequently processed for change information extraction through the application of the U-shaped architecture and the LMGM. The U-shaped architecture leverages the Transformer's deep features to guide CNN shallow features in obtaining change-related features. Simultaneously, CNN features are employed to complement fine-grained details. This synthesis effectively transforms multiscale architectural features into accurate change features. On the other hand, the LMGM iteratively refines and reconstructs multiscale change features, ultimately producing high-quality change maps. Furthermore, the incorporation of the CFMAE significantly expedites network convergence, enhances generalization capabilities, and unlocks the full potential of the Transformer structure. The proposed U-Conformer is extensively evaluated on challenging datasets: LEVIR-CD (Google Earth), WHU-CD (Aerial), and GZ-CD (Google Earth), each presenting unique difficulties. A series of ablation experiments further corroborate the effectiveness of U-Conformer. Finally, comprehensive comparisons with ten other SOTA methods affirm that U-Conformer outperforms its counterparts, delivering superior performance.

In our future research, we will focus on the following three aspects.

1) We will further expand the strategy of combining ViTs and CNNs to integrate local and global information more effectively.
2) We will continue to enlarge the dataset for unsupervised pretraining to match the capabilities of a more powerful backbone network.
3) We intend to explore the development of a unified model that extends our approach to all CD tasks.

REFERENCES

[1] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.
[2] G. Chen, G. J. Hay, L. M. T. Carvalho, and M. A. Wulder, "Object-based change detection," *Int. J. Remote Sens.*, vol. 33, no. 14, pp. 4434–4457, 2012.

[3] R. Manonmani and G. Suganya, "Remote sensing and GIS application in change detection study in urban zone using multi temporal satellite," *Int. J. Geomatics Geosci.*, vol. 1, no. 1, pp. 60–65, 2010.

[4] R. K. Gupta, "Change detection techniques for monitoring spatial urban growth of Jaipur city," *Inst. Town Planners India J.*, vol. 8, no. 3, pp. 88–104, 2011.

[5] C. A. Mucher, K. T. Steinnocher, F. P. Kressler, and C. Heunks, "Land cover characterization and change detection for environmental monitoring of pan- Europe," *Int. J. Remote Sens.*, vol. 21, no. 6/7, pp. 1159–1181, 2000.

[6] K. S. Willis, "Remote sensing change detection for ecological monitoring in United States protected areas," *Biol. Conservation*, vol. 182, pp. 233–242, 2015.

[7] B. Gokaraju, A. C. Turlapaty, D. A. Doss, R. L. King, and N. H. Younan, "Change detection analysis of tornado disaster using conditional copulas and data fusion for cost-effective disaster management," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 2015, pp. 1–8.

[8] H. Qiao, X. Wan, Y. Wan, S. Li, and W. Zhang, "A. novel change detection method for natural disaster detection and segmentation from video sequence," *Sensors*, vol. 20, 2020, Art. no. 5076.

[9] A. Agapiou, "UNESCO world heritage properties in changing and dynamic environments: Change detection methods using optical and radar satellite data," *Heritage Sci.*, vol. 9, no. 1, pp. 1–14, 2021.

[10] N. Lercari, "Monitoring earthen archaeological heritage using multitemporal terrestrial laser scanning and surface change detection," *J. Cultural Heritage*, vol. 39, pp. 152–165, 2019.

[11] A. C. Mondini, F. Guzzetti, P. Reichenbach, M. Rossi, M. Cardinali, and F. Ardizzone, "Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images," *Remote Sens. Environ.*, vol. 115, no. 7, pp. 1743–1757, 2011.

[12] K. He et al., "Transformers in medical image analysis," *Intell. Med.*, vol. 3, no. 1, pp. 59–78, 2023.

[13] E. F. Lambin and A. H. Strahlers, "Change-vector analysis in multitemporal space: A tool to detect and categorize land-cover change processes using high temporal-resolution satellite data," *Remote Sens. Environ.*, vol. 48, no. 2, pp. 231–244, 1994.

[14] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using tri-temporal logic-verified change vector analysis," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 278–293, 2020.

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

[16] M. H. Kesikoğlu, Ü. H. Atasever, and C. Özkan, "Unsupervised change detection in satellite images using fuzzy C-means clustering and principal component analysis. The international archives of the photogrammetry," *Remote Sens. Spatial Inf. Sci.*, vol. 40, pp. 129–132, 2013.

[17] S. Marchesi and L. Bruzzone, "ICA and kernel ICA for change detection in multispectral remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2009, vol. 2, pp. II-980–II-983.

[18] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

[19] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998.

[20] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[21] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021, doi: 10.1109/JSTARS.2020.3037893.

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

[23] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012, doi: 10.1109/TGRS.2022.3207551.

[24] Z. Lv, J. Liu, W. Sun, T. Lei, J. A. Benediktsson, and X. Jia, "Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411115, doi: 10.1109/TGRS.2023.3334521.

[25] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015, doi: 10.1109/TGRS.2023.3241257.

[26] F. Zhou, C. Xu, R. Hang, R. Zhang, and Q. Liu, "Mining joint intra-and inter-image context for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4403712, doi: 10.1109/TGRS.2023.3275819.

[27] Z. Lv, H. Huang, W. Sun, T. Lei, J. A. Benediktsson, and J. H. Li, "Novel enhanced UNet for change detection using multimodal remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2505405, doi: 10.1109/LGRS.2023.3325439.

[28] R. Hang, S. Xu, P. Yuan, and Q. Liu, "AANet: An ambiguity-aware network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art no. 5612911, doi: 10.1109/TGRS.2024.3371463.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[30] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[31] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[32] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.

[33] W. G. Chaminda Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[34] Y. Li, L. Wang, W. Mi, H. Xu, J. Hu, and H. Li, "Distracted driving detection by combining VIT and CNN," in *Proc. IEEE 25th Int. Conf. Comput. Supported Cooperative Work Des.*, 2022, pp. 908–913.

[35] S. Yu, M. He, R. Nie, C. Wang, and X. Wang, "An unsupervised hybrid model based on CNN and VIT for multimodal medical image fusion," in *Proc. 2nd Int. Conf. Electron., Commun. Inf. Technol.*, 2021, pp. 235–240.

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

[37] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8000605.

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

[39] M. Kumar, D. K. Yadav, S. Ray, and R. Tanwar, "Handling illumination variation for motion detection in video through intelligent method: An application for smart surveillance system," *Multimedia Tools Appl.*, vol. 83, pp. 29139–29157, 2023.

[40] L. He, S. Jiang, X. Liang, N. Wang, and S. Song, "Diff-Net: Image feature difference based high-definition map change detection for autonomous driving," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 2635–2641.

[41] M. K. Panda, A. Sharma, V. Bajpai, B. N. Subudhi, V. K. Thangaraj, and V. Jakhetiya, "Encoder and decoder network with resnet-50 and global average feature pooling for local change detection," *Comput. Vis. Image Understanding*, vol. 222, 2022, Art. no. 103501.

[42] X. Li, H. Duan, J. Li, Y. Deng, and F.-Y. Wang, "Biological eagle eye-based method for change detection in water scenes," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108203.

[43] X. Zhao, G. Wang, Z. He, and H. Jiang, "A survey of moving object detection methods: A practical perspective," *Neurocomputing*, vol. 503, pp. 28–48, 2022.

[44] D. Xia, J. Huang, J. Yang, X. Liu, and H. Wang, "DuARUS: Automatic Geo-object change detection with street-view imagery for updating road database at Baidu maps," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 3565–3574.

[45] M. Biparva, D. Fernández-Llorca, R. I. Gonzalo, and J. K. Tsotsos, "Video action recognition for lane-change classification and prediction of surrounding vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 569–578, Sep. 2022.

[46] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auton. Robots*, vol. 42, pp. 1301–1322, 2018.

[47] Z. Wang, Y. Zhang, L. Luo, and N. Wang, "TransCD: Scene change detection via transformer-based architecture," *Opt. Exp.*, vol. 29, no. 25, pp. 41409–41427, 2021.
[48] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 745–746, Aug. 2000.
[49] I. Haritaoglu, D. Harwood, and L. S. Davis, "E 4S: A real-time system for detecting and tracking people in 2 1/2D," in *Proc. 5th Eur. Conf. Comput. Vis.*, 1998 pp. 877–892.
[50] R. O'Callaghan and T. Haga, "Robust change-detection by normalised gradient-correlation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[51] X. Liu, Y. Wu, W. Liang, Y. Cao, and M. Li, "High resolution SAR image classification using global-local network structure based on vision transformer and CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4505405, doi: 10.1109/LGRS.2022.3151353.
[52] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 913–919.
[53] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim, "Enriched CNN-transformer feature aggregation networks for super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4956–4965.
[54] H. Yang, H. Yu, K. Zheng, J. Hu, T. Tao, and Q. Zhang, "Hyperspectral image classification based on interactive transformer and CNN with multilevel feature fusion network," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5507905, doi: 10.1109/LGRS.2023.3303008.
[55] Y. Xia, Y. Xiong, and K. Wang, "A transformer model blended with CNN and denoising autoencoder for inter-patient ECG arrhythmia classification," *Biomed. Signal Process. Control*, vol. 86, 2023, Art. no. 105271.
[56] X. Chen, D. Li, M. Liu, and J. Jia, "CNN and transformer fusion for remote sensing image semantic segmentation," *Remote Sens.*, vol. 15, no. 18, p 4455, 2023, doi: 10.3390/rs15184455.
[57] J. Zhang, "Towards a high-performance object detector: Insights from drone detection using ViT and CNN-based deep learning models," in *Proc. IEEE Int. Conf. Sensors, Electron. Comput. Eng.*, 2023, pp. 141–147.
[58] H. Du et al., "Bilateral semantic fusion siamese network for change detection from multitemporal optical remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6003405, doi: 10.1109/LGRS.2021.3082630.
[59] 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, doi: 10.1109/LGRS.2021.3056416.
[60] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9414–9423.

Yongjing Cui (Student Member, IEEE) was born in Chongqing, China, in 1999. He received the B.S. degree in electronics and information science and technology from Sichuan University, Sichuan, China, in 2021. He is currently working toward the M.S. degree in electronic and information engineering with the Beijing Institute of Technology, Beijing, China.

His research interests include remote sensing semantic segmentation and change detection.

He Chen (Member, IEEE) was born in Shenyang, China in 1970. She received the Ph.D. degree in electronic engineering from the Harbin Institute of Technology, Harbin, China, in the 1998.

She is currently a Professor and Dean with the School of Information, Beijing Institute of Technology, Beijing, China. Her research interests include system-on-chip design, remote sensing data intelligent processing, and very-large-scale integration architectures for real-time image and signal processing.

Shan Dong (Member, IEEE) was born in Hebei, China, in 1992. She received the B.S. degree in electronic information engineering from the Dalian University of Technology, Dalian, China, in 2014, the M.S. degree in electronics and communications engineering from the Beijing Institute of Technology, Beijing, China, in 2016, and the Ph.D. degree in signal and information processing from the Communication University of China, Beijing, in 2022.

She is currently an Postdoctoral Researcher with the Beijing Institute of Technology. Her research interests include remote sensing semantic segmentation and change detection.

Guanqun Wang (Member, IEEE) was born in Heilongjiang, China, in 1995. He received the B.S. degree in electronics and communications engineering and the Ph.D. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China, in 2017 and 2013, respectively.

He is currently a Postdoctoral Researcher with the School of Computer Science, Peking University, Beijing. His research interests include remote sensing object detection, recognition, and multimodal large language models.

Yin Zhuang (Member, IEEE) was born in Henan, China, in 1990. He received the B.S. degree in electronic engineering from the University of Sussex, Brighton, U.K., in 2013, and the Ph.D. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China, in 2018.

From 2018 to 2020, he was a Postdoc with the School of Electronic Engineering and Computer Sciences, Peking University, Beijing. Since January 2021, he has been a tenure-track Assistant Professor with the School of Information and Communication Engineering, Beijing Institute of Technology. His research interests include remote sensing object detection and recognition