From Macro to Micro: A Lightweight Interleaved Network for Remote Sensing Image Change Detection

Yetong Xu, Tao Lei, *Senior Member, IEEE*, Hailong Ning, Shaoxiong Lin, Tongfei Liu, *Member, IEEE*, Maoguo Gong, *Fellow, IEEE*, and Asoke K. Nandi, *Life Fellow, IEEE*

***Abstract*— Remote sensing image change detection (RSICD) is a crucial technology for Earth monitoring, but it faces two major challenges in practical applications. First, the complex scenes in remote sensing (RS) images make it difficult to accurately locate and distinguish the small change targets. Second, most RSICD methods usually stack complex modules to improve model performance, which inevitably leads to high computational costs. To address these issues, this article proposes a lightweight interleaved network from macro to micro for RSICD, named M2M-LINet. First, a three-stage RSICD framework inspired by holistic registration theory in biological vision is designed to deal with complex and small change targets. Specifically, the first stage is the coarse location stage to obtain a coarse localization on change targets, which is similar to the macroscopic observation of human eyes. The second stage is the fine detail focusing stage for capturing fine-grained information by designing a CNN–Transformer interleaved model, which is akin to the microscopic observation of the human eye. The third stage is the decoding prediction stage for predicting the final change features by mimicking the human interpretation of change information. Second, a lightweight interleaved structure is designed in the fine detail focusing stage to reduce the model size for facilitating deployment. It consists of two critical lightweight components. One is the lightweight convolutional block (LCB) that is devised to enhance high-frequency and low-frequency information for fine-grained change localization. The other is the lightweight Transformer block (LTB) that is designed as a linear self-attention mechanism to integrate multiscale information into feature maps effectively. The experiments on three RSICD datasets demonstrate that the proposed method performs excellently for complex change scenes and small change targets and significantly reduces the consumption of computing resources.**

***Index Terms*— Change detection (CD), lightweight, remote sensing (RS) image, Transformer.**

Received 18 August 2024; revised 18 December 2024 and 19 February 2025; accepted 3 March 2025. Date of publication 5 March 2025; date of current version 24 March 2025. This work was supported in part by the National Natural Science Foundation of China Program under Grant 62271296, Grant 62201452, Grant 62201334, and Grant 62301302; and in part by the Scientific Research Program Funded by Shaanxi Provincial Education Department under Grant 23JP014 and Grant 23JP022. *(Corresponding author: Tao Lei.)*

Yetong Xu, Tao Lei, Shaoxiong Lin, and Tongfei Liu are with Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China. (e-mail: xuyetong1999@163.com; leitao@sust.edu.cn; 231611039@sust.edu.cn; liutongfei_home@hotmail.com).

Hailong Ning is with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: ninghailong93@gmail.com).

Maoguo Gong is with the Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, Xian 710071, P. R. China (e-mail: gong@ieee.org).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University of London, UB8 3PH Uxbridge, U.K. (e-mail: asoke.nandi@brunel.ac.uk).

Digital Object Identifier 10.1109/TGRS.2025.3548562

I. Introduction

REMOTE sensing image change detection (RSICD) aims to identify surface changes by analyzing multitemporal images of the same geographical location. It has been widely used in various fields, including disaster assessment [1], urban expansion [2], land planning [3], and agricultural monitoring [4]. In recent years, the applications of change detection (CD) techniques have become more and more important due to the deterioration of the natural environment and rapid urbanization. Although RSICD has achieved great success, it still faces two main challenges. *1) Impact of interference factors,* such as illumination variations and seasonal changes, can cause pseudo-changes in images. Unrelated dynamic changes may also affect detection: *2) unbalanced change targets. There are clear size differences between changed and unchanged areas, causing class imbalance. Moreover, changed regions vary in size and shape, which increases the difficulty of localization and completeness detection.* To address the above challenges, many RSICD studies have been proposed, which can be mainly divided into three categories: CNN-based methods, Transformer-based methods, and CNN–Transformer hybrid methods.

Convolutional neural networks (CNNs) are widely used in RSICD tasks [5], [6], [7] due to their excellent feature extraction abilities. Because of the bi-temporal characteristic of RSICD, Siamese CNN methods have become the preferred architecture. Furthermore, deep Siamese CD networks with additional stacked convolutional blocks [8], [9], [10] have been proposed to capture deeper semantic representations and improve RSICD performance. Various attention mechanisms [11], [12], [13] have also been introduced into these networks for finer feature extraction. Additionally, multiscale fusion strategies [14], [15], [16] have been employed to aggregate multilevel information and enhance the semantic representation of bi-temporal images. However, these CNN-based methods still struggle to model global information

of bi-temporal images, which significantly limits their performance.

Transformers overcome the global modeling limitations of CNNs and have demonstrated excellent performance in many computer vision tasks [17], [18], [19], including the advancements in RSICD. Transformer-based methods employ a Siamese Transformer encoder–decoder architecture [20], [21], [22], [23], [24] to model bi-temporal images, effectively capturing global change information. Additionally, channel-based self-attention mechanisms [25], [26] are utilized to capture spectral information between bi-temporal images, improving the accuracy and robustness of feature extraction. However, despite their excellent performance, Transformer-based methods suffer from high computational complexity and exhibit limited ability to extract local information, leading to insufficient attention to image details.

To solve the limitations of the aforementioned methods, the hybrid networks of CNNs and Transformers [27], [28], [29], [30], [31] are designed. They can simultaneously acquire local and global features of RS images and have become the mainstream paradigm of RSICD tasks. Hybrid architecture is mainly classified into serial and parallel. Specifically, the serial architecture can be further divided into two groups: 1) CNNs followed by Transformers [22], [25] [Fig. 1(a)]: the CNN first extracts shallow information and then passes them to Transformers to capture global information, improving global modeling efficiency and 2) Transformers followed by CNNs [21] [Fig. 1(b)]: The Transformer first captures global features, and then CNN enhances local details to adapt to complex scenarios. The parallel architecture [28], [29] [Fig. 1(c)]: CNNs and Transformers extract features in parallel and then perform interaction operations on the local features from the CNN and the global features from the Transformer. While the serial architecture processes image features in different stages, it limits interaction between features, making it difficult to detect complex changes accurately. In contrast, the parallel architecture improves the information interaction to some extent but still faces two major challenges: 1) small change targets in complex scenes are still difficult to accurately detect and 2) the hybrid dual-branch architecture introduces a large number of parameters and a high computational cost.

To tackle the above challenges, we propose a novel lightweight interleaved network for RSICD. This method is inspired by the holistic registration theory [32] in biological vision (Navon, 1977), which emphasizes that human visual cognition tends to first recognize global features and then focus on local details. Intuitively, this is similar to the cognitive pattern of "seeing the forest before the trees." First, the three-stage architecture progressively learns and refines change details, effectively improving the detection accuracy of small change targets in complex scenes. Second, lightweight CNN and Transformer blocks are alternately designed in this interleaved structure. This design not only improves the efficiency of information exchange but also reduces the parameters and computational complexity of the model. The main contributions of this article can be summarized as follows.

1) A novel three-stage RSICD architecture is proposed to address the problems of small change targets in complex scenes. Distinct from the mainstream encoder–decoder network, our architecture has three critical stages: the coarse location stage captures overall change trends and approximate locations, the fine detail focusing stage refines detailed and accurate change information, and the decoding prediction stage predicts the final change features with boundary assistance.
2) A new lightweight CNN–Transformer interleaved structure is designed to reduce computational consumption and refine features. Unlike the existing hybrid architectures, the lightweight LCB and LTB enhance local and global features. LCB enhances high-frequency and low-frequency local information, while LTB employs a linear self-attention mechanism for efficient global modeling.
3) Based on the above design, a three-stage lightweight CNN–Transformer interleaved network for RSICD tasks is proposed. It gradually improves change information from coarse to fine, effectively avoiding interference factors in RSICD. This method not only improves the accuracy and completeness of RSICD but also minimizes the model size and computational complexity.

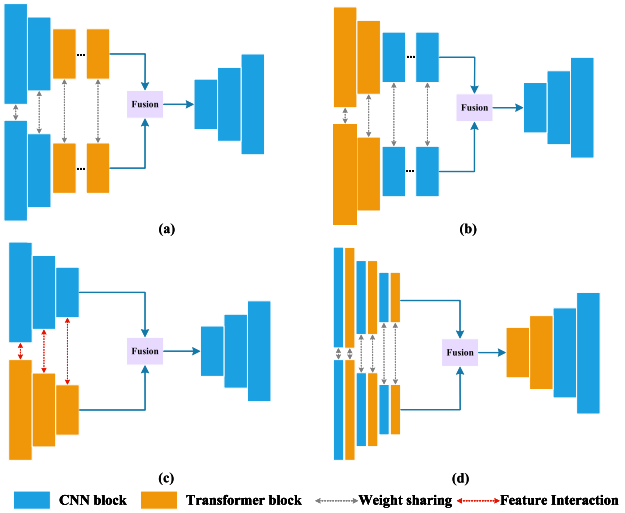

Fig. 1. Comparison of different types of hybrid structures. (a) Serial: the CNN followed by the Transformer. (b) Serial: the Transformer followed by the CNN. (c) Parallel: the CNN and the Transformer dual-branch. (d) Interleaved: the CNN and the Transformer.

II. Related Works

A. Traditional Methods

Traditional RSICD methods are mainly divided into pixel-based, feature-based, and object-based methods. Pixel-based RSICD methods identify change regions by analyzing spectral differences between individual pixels. Common methods include pixel classification based on threshold segmentation [33] and change detection based on clustering algorithms [34]. However, these methods heavily depend on image content and preprocessing quality and are easily affected by noise. In contrast, feature-based RSICD methods focus on feature extraction and classifier design. Common methods include principal component analysis (PCA) [35] and Gabor filters [36]. However, they are sensitive to seasonal changes,

atmospheric conditions, and noise, which reduces detection accuracy and stability. Object-based RSICD methods use objects as the detection unit [37], [38], combining various image information such as shape, texture, and spectral data, offering higher accuracy and robustness compared to pixel-based methods. However, these traditional methods require high-quality images, and the features are susceptible to seasonal and sensor differences. Additionally, they rely heavily on prior knowledge, limiting the model's generalization ability.

B. CNN-Based Methods

CNNs have achieved remarkable results in many fields and are widely used in end-to-end RSICD tasks. According to the characteristics of RSICD tasks, Daudt et al. [5] first applied Siamese fully convolutional networks (FCNs) and proposed three end-to-end architectures. However, the simple structure of the Siamese network inherently limits its representation ability, leading to suboptimal performance for complex scenes. To address the issue, various improvement strategies have been introduced, including attention mechanisms, multiscale feature fusion modules, and deeply supervised strategies, yielding satisfactory results. Attention mechanisms can significantly improve the feature representation ability of networks for RSICD. For example, MAFGNet [39] employs the channel-attention mechanism employs the channel-attention mechanism to select important spectral channels dynamically. AGCDetNet [40] applies the spatial-attention mechanism to focus on specific regions and thus enhance their spatial information feature representation for RSICD tasks. DESSNet [14] employs the nonlocal attention mechanism to extract global information to overcome narrow contextual receptive fields. By combining the various attention mechanisms [41], [42], [43], the improved networks can provide better feature representation to achieve accurate RSICD. Additionally, the multiscale feature fusion strategies aggregate features from different scales, enabling models to focus on both semantic and detailed changes. For example, MSDFFN [44] and DGMA2-Net [45] apply a multiscale fusion strategy to combine features from different scales to improve the accuracy of RSICD results. Furthermore, deeply supervised strategies help models learn features at various levels. For instance, DSAMNet [11] and A2Net [15] introduce deeply supervised signals at multiple levels of the network, which significantly improved the performance of CNN-based methods. Despite these advancements, CNN-based methods still struggle to effectively capture global information from bi-temporal images, which limits their performance.

C. Transformer-Based Methods

Recently, Transformers have been successfully applied to remote sensing (RS) image interpretation tasks due to their excellent ability to model long-range dependency. The prevailing Transformer-based methods mainly utilize standard Vision Transformer (ViT) [46] or Swin Transformer [47] to construct RSICD networks. Among the standard ViT-based methods, BIT [20] is the pioneering work of applying the Transformer to RSICD tasks. It captures important features with semantic tokens and introduces a dual-branch Transformer encoder–decoder for efficient encoding. ForestViT [48] leverages the benefits of the self-attention mechanism in ViT to design a multilabel classification Vision Transformer model specifically for deforestation detection. CSANet [26] also adopts a Siamese Transformer and enhanced feature representation with spatial and spectral attention. However, the computational complexity of self-attention in ViT scales quadratically with image size. To address this issue, the Swin Transformer introduces a window partitioning mechanism to optimize self-attention computation. Among Swin Transformer-based methods, SwinSUNet [23] proposes a pure Swin Transformers network with a Siamese U-shape structure to capture global context information. TransY-Net [24] uses the Swin Transformer to learn discriminative global features and aggregates multilevel features from a pyramid structure to enhance feature representation. Furthermore, M-Swin [4] designs a Siamese Swin Transformer with hierarchical windows, effectively capturing change information for small targets. It provides clear target boundaries and improved RSICD accuracy by fusing multiscale features from multiple windows.

D. CNN–Transformer Hybrid Networks

Practically, neither CNNs nor Transformers can achieve both local and global modeling. To combine the strengths of CNNs and Transformers, various hybrid architectures have been proposed. For example, ChangeFromer [21] designs a serial structure with a Transformer followed by a CNN. The Transformer captures long-range dependency while the CNN aggregates multiscale local features. TransCD [49] and AMTNet [50] employ a serial structure with a CNN followed by a Transformer. The CNN is first used to extract multilevel local features and then the Transformer module is used to model the contextual information of bi-temporal images. Besides, ICIF [28] and WNet [29] utilize a parallel structure of dual-branch with the CNN and the Transformer. They capture features through interactions to learn diverse local and global features. Additionally, some complex hybrid methods have been proposed to combine the CNN and the Transformer in a single module. For example, ACAHNet [27] introduces asymmetric multihead crossover attention by combining the advantages of the CNN and the Transformer. LSAT [51] proposes a cross-dimensional interactive self-attention module that combines the CNN with self-attention across the channel and spatial dimensions to obtain richer multidimensional features.

Despite significant advances in RSICD, the unique challenges in RSICD tasks are not fully explored, resulting in suboptimal performance when existing methods confront complex scenes and small change targets. Besides, most of the current methods also suffer from a large number of network parameters and slow inference speed. To tackle the above challenges, we propose a three-stage lightweight interleaved method for RSICD tasks. This method not only achieves high detection accuracy, especially for complex scenes and small change targets but also effectively reduces the number of model parameters and computational costs.

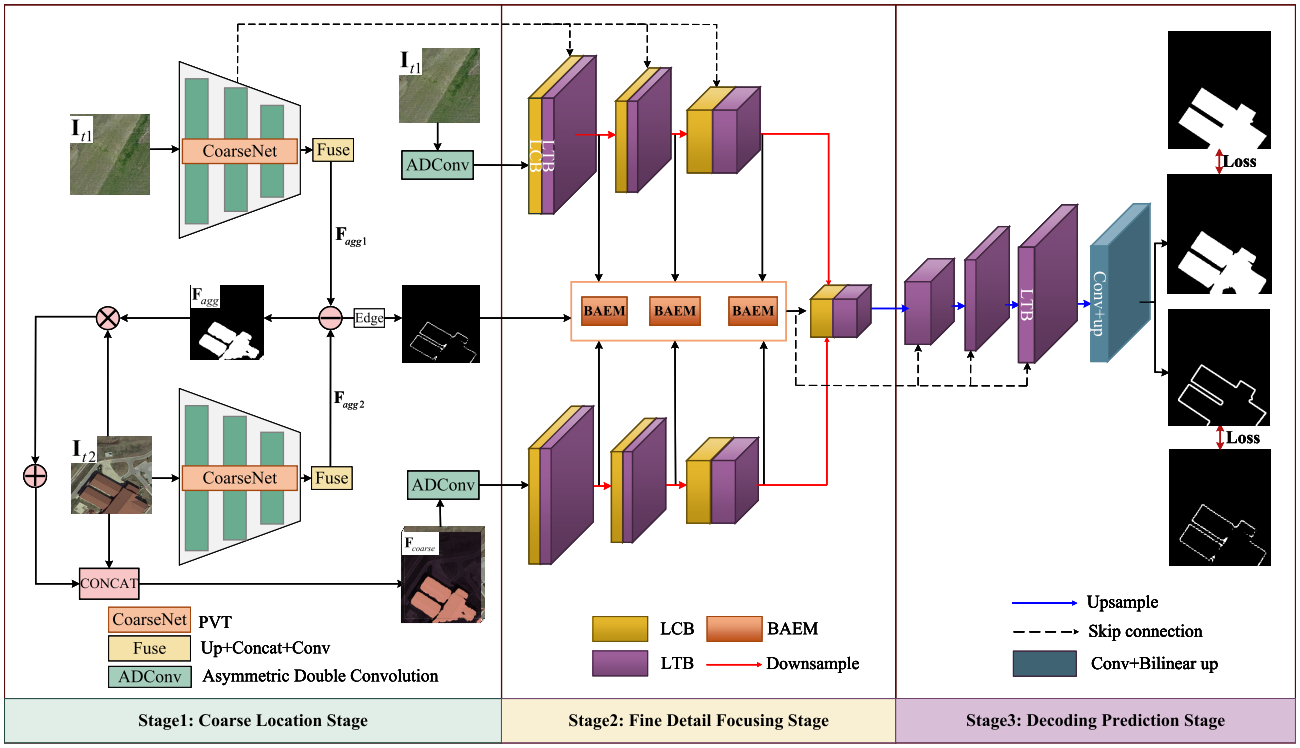

Fig. 2. Architecture of the proposed M2M-LINet. First, the coarse location stage extracts feature maps from the input bi-temporal images. Next, these features are sent to the fine detail focusing stage to generate enhanced bi-temporal features. Finally, pixel-level change prediction is achieved during the decoding prediction stage.

III. Methodology

A. Overall Framework

The proposed three-stage lightweight interleaved network M2M-LINet is shown in Fig. 2. First, in the coarse location stage, the coarse aggregation localization module (CALM) is devised to extract coarse features from bi-temporal images for preliminary change localization. Second, the coarse features are input into the fine detail focusing stage for generating enhanced bi-temporal features by the interleaved architecture of the CNN and the Transformer. Finally, in the decoding prediction stage, the boundary-aware enhancement module (BAEM) serves as an auxiliary branch to optimize the decoding process using skip connections. Specifically, BAEM integrates boundary information to enhance the post-temporal image, significantly improving the details of changed targets. In addition, to reduce the model complexity and the number of network parameters, the M2M-LINet model includes two critical lightweight components: 1) the LCB is presented to enhance both high-frequency and low-frequency information for improving change localization using lightweight convolution and 2) the LTB is presented to effectively merge multiscale information for achieving accurate RSICD using a lightweight agent attention mechanism.

B. Three-Stage CD Architecture

1) Coarse Location Stage: In the coarse location stage, the CALM is designed for efficient and accurate RSICD by integrating multiscale features, as shown in Fig. 2. The specific steps are as follows. First, a pretrained pyramid Visual Transformer (PVT) [52] is employed as the coarse network to extract multiscale features $\mathbf{F}_L (L \in \{1, 2, 3\})$ which is crucial for locating the changed regions. Then, the extracted three-layer features are upsampled layer by layer and concatenated with the neighboring features to obtain aggregated features $\mathbf{F}_{\text{agg1}}$ and $\mathbf{F}_{\text{agg2}}$ through convolution aggregation

$$\mathbf{F}_{\text{agg}i} = C_3\left[\text{Up}(C_3\left[\text{Up}(\mathbf{F}_1), \mathbf{F}_2\right]), \mathbf{F}_3\right] \tag{1}$$

where Up denotes the upsampling, [,] is the concatenation operation, and C_3 is a 3×3 convolution operation, $i = 1, 2$.

Next, an aggregated change feature $\mathbf{F}_{\text{agg}}$ is calculated by (2) to obtain the difference information between $\mathbf{F}_{\text{agg1}}$ and $\mathbf{F}_{\text{agg2}}$

$$\mathbf{F}_{\text{agg}} = \mathbf{F}_{\text{agg1}} - \mathbf{F}_{\text{agg2}}. \tag{2}$$

Then, the aggregated change feature $\mathbf{F}_{\text{agg}}$ is multiplied with the image $\mathbf{I}_{t2}$ and combined with the original image $\mathbf{I}_{t2}$ to obtain the final coarse-grained location feature map $\mathbf{F}_{\text{coarse}}$. Finally, $\mathbf{F}_{\text{coarse}}$ is mapped to a high-dimensional feature space by ADConv for subsequent processing with the fine detail focusing stage. This stage can be expressed as

$$\mathbf{F}_{\text{coarse}} = \text{ADConv}\left(\left[\left(\left(\mathbf{F}_{\text{agg}} \odot \mathbf{I}_{t2}\right) \oplus \mathbf{I}_{t2}\right), \mathbf{I}_{t2}\right]\right) \tag{3}$$

where [,] is the concatenation operation and ADConv [53] is the asymmetric double convolution.

2) Fine Detail Focusing Stage: In the fine detail focusing stage, the coarse-grained features obtained in the coarse location stage are refined and enhanced more precisely. Specifically, to obtain richer details, a novel interleaved architecture combining CNNs and Transformers is introduced. Unlike

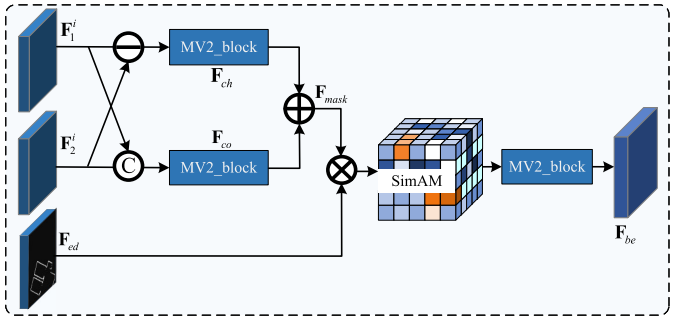

Fig. 3. Illustration of our BAEM.

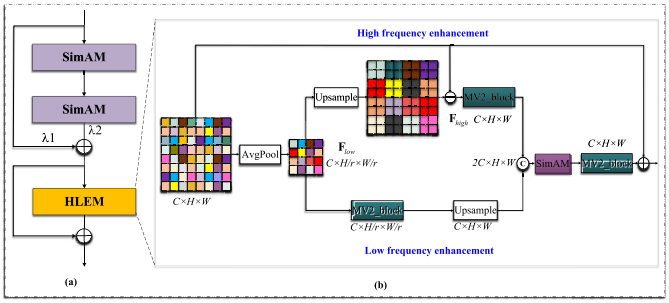

Fig. 4. Illustration of our LCB. (a) LCB. (b) HLEM is designed to enhance high-frequency and low-frequency information.

traditional serial or parallel architectures, the interleaved architecture integrates CNN and Transformer components alternately at each stage, facilitating comprehensive interaction between local and global information. Moreover, the dual-branch interleaved architecture employs weight sharing without introducing extra parameters. This design not only effectively improves information interaction, but also significantly reduces the parameter number and computational complexity of the model. Specifically, the architecture includes two crucial lightweight modules: LCB and LTB.

The features obtained from the coarse location are first processed by the LCB to enhance both high-frequency and low-frequency information of images. After enhancement by LCB, the features are further processed by LTB which captures global information and integrates it with local features. The interleaved architecture uses LCB and LTB alternately and combines the advantages of CNNs in extracting image details and Transformers in capturing image global information. It maintains a lightweight structure while accurately identifying image detail changes.

Detailed module design and implementation are discussed in Section III-C. Experimental results show that this interleaved architecture performs well in complex scenes, significantly improving the accuracy and efficiency of RSICD.

3) Decoding Prediction Stage: In the decoding prediction stage, the LTB and a convolutional prediction head are employed for final predictions. Additionally, because the boundaries of targets are crucial for RSICD, a boundary-aware enhancement module (BAEM) is designed as an auxiliary branch to enhance boundary features during the decoding stage. As shown in Fig. 3, the BAEM consists of three branches, and each of them processes features from different inputs. Concretely, the three inputs are the output features from each layer of the Siamese encoder, that is, $\mathbf{F}_1^i$, $\mathbf{F}_2^i$, and the boundary feature map $\mathbf{F}_{\text{ed}}$. The first branch of the BAEM captures and enhances the change information $\mathbf{F}_{\text{ch}}$ based on the difference between $\mathbf{F}_1^i$ and $\mathbf{F}_2^i$. It highlights the changes between the bi-temporal images to help the model identify and distinguish changed areas. The second branch concatenates two encoded features to obtain the common information $\mathbf{F}_{\text{co}}$

$$\mathbf{F}_{\text{ch}} = \text{MV2}\left(\mathbf{F}_1^i \ominus \mathbf{F}_2^i\right), \quad \mathbf{F}_{\text{co}} = \text{MV2}\left(\left[\mathbf{F}_1^i, \mathbf{F}_2^i\right]\right) \tag{4}$$

where $[,]$ is the concatenation operation and MV2 is the inverted residual block in MobileNetv2 [54]. Next, the change information $\mathbf{F}_{\text{ch}}$ is integrated with the common features $\mathbf{F}_{\text{co}}$ to obtain $\mathbf{F}_{\text{mask}}$

$$\mathbf{F}_{\text{mask}} = \mathbf{F}_{\text{ch}} \oplus \mathbf{F}_{\text{co}}. \tag{5}$$

In the third branch, the boundary feature map $\mathbf{F}_{\text{ed}}$ is utilized to further refine and optimize the enhanced change information. The boundary feature map aids in capturing target contours and details of images, ensuring that the final output $\mathbf{F}_{\text{be}}$ has clear and accurate contours

$$\mathbf{F}_{\text{be}} = \text{MV2}(\text{SimAM}(\mathbf{F}_{\text{ed}} \otimes \mathbf{F}_{\text{mask}}) \oplus \mathbf{F}_{\text{mask}})) \tag{6}$$

where SimAM [55] is a lightweight 3-D attention block.

Through the collaboration of the three branches, the BAEM module effectively enhances and optimizes the boundary of changed targets, resulting in more accurate RSICD results.

C. Lightweight Interleaved Structure

1) Lightweight Convolutional Block: To effectively enhance RSICD performance, an improved LCB is proposed by integrating attention mechanisms and spatial enhancement techniques, as shown in Fig. 4. Specifically, in the LCB module, information from the previous layer is initially processed using two lightweight and efficient attention modules, namely SimAM. Then adaptive weighting using λ_1 and λ_2 allows for more variable information to improve sensitivity for changed regions.

To further enhance feature representation, a high-frequency and low-frequency enhancement module (HLEM) is proposed. Due to the complexity of obtaining frequency domain information and the difficulty of embedding Fourier transforms into CNNs, we present a differentiable spatial domain enhancement method. It effectively separates and enhances high-frequency and low-frequency information in RS images. Specifically, a pooling operation is first applied to downsample feature maps, yielding simple and effective low-frequency features $\mathbf{F}_{\text{low}}$. Subsequently, the high-frequency information $\mathbf{F}_{\text{high}}$ is obtained by subtracting the upsampled low-frequency feature from the original feature $\mathbf{F}_{\text{original}}$ as follows:

$$\mathbf{F}_{\text{high}} = \mathbf{F}_{\text{original}} - \text{Up}(\mathbf{F}_{\text{low}}). \tag{7}$$

After obtaining the high-frequency and low-frequency features, convolutional enhancement is performed separately to obtain the enhanced high-frequency feature $\mathbf{F}_{\text{high\_en}}$ and the enhanced low-frequency feature $\mathbf{F}_{\text{low\_en}}$. One part focuses on

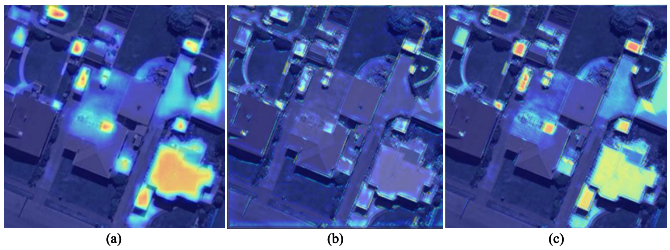

Fig. 5. Visualization heatmap of the HLEM module. (a) Low-frequency heat map. (b) High-frequency heat map. (c) Frequency enhancement heat map.

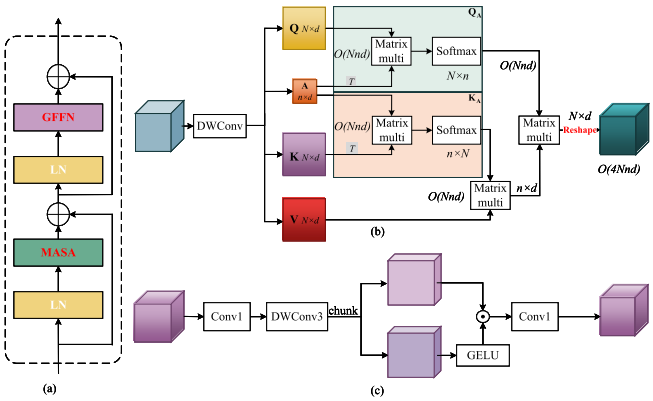

Fig. 6. Illustration of our LTB. (a) LTB. (b) MASA. (c) GFFN.

the high-frequency details of images, while the other concentrates on the low-frequency content and semantics as follows:

$$\mathbf{F}_{\text{high\_en}} = \text{MV}_2\left(\mathbf{F}_{\text{high}}\right) \tag{8}$$

$$\mathbf{F}_{\text{low\_en}} = \text{MV}_2(\mathbf{F}_{\text{low}}). \tag{9}$$

Next, the enhanced high-frequency and low-frequency features are further combined. First, the low-frequency information is upsampled to align with the dimensions of the high-frequency features. Second, they are concatenated and fused, resulting in the final enhanced feature $\mathbf{F}_{\text{ce}}$ as follows:

$$\mathbf{F}_{\text{ce}} = \text{MV}_2(\text{SimAM}([\text{Up}(\mathbf{F}_{\text{low\_en}}), \mathbf{F}_{\text{high\_en}}])). \tag{10}$$

Fig. 5 is a visualized heatmap, indicating that the module effectively enhances the high-frequency and low-frequency information of the image, thereby improving the effectiveness of image processing results.

2) Lightweight Transformer Block: LTB consists of two components: multihead agent self-attention module (MASA) and gated feed-forward network (GFFN), as shown in Fig. 6. Concretely, the semantic agent attention in MASA effectively combines linear attention (LA) and traditional self-attention (SA). It reduces computational complexity and has excellent global information extraction capability. GFFN is a novel gated feed-forward network. It introduces the gating mechanism to suppress invalid and redundant features, allowing only discriminative information to pass through the network layer. This design ensures controlled transformation of features, enhancing both the efficiency and effectiveness of the model.

Previous studies have proposed several optimization strategies to mitigate the quadratic increase in computational complexity associated with self-attention. For example, window-based attention [47], [56] reduces the computational cost by restricting attention calculation within a local window range, and sparse attention [57], [58] reduces the computational cost by reducing the number of key features ($\mathbf{K}$, $\mathbf{V}$). More recently, the LA [59] reduces computational complexity to $O(N)$ by changing the sequence of matrix multiplication. Though these methods are effective, they still struggle with the issue of the limited expressive power of LA.

In contrast, agent self-attention extends the traditional triplet tensor $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ into a quadruple $\mathbf{Q}, \mathbf{A}, \mathbf{K}, \mathbf{V}$ by introducing a representative semantic additional tensor $\mathbf{A}$. Concretely, the low dimensional agent $\mathbf{A}$ is first used to aggregate information from $\mathbf{K}$ and $\mathbf{V}$, and then $\mathbf{A}$ is employed to aggregate $\mathbf{Q}$ to obtain $\mathbf{Q}_{\mathbf{A}}$ for the final attention calculation. Since the dimensions of the agent $\mathbf{A}$ are much smaller than the query $\mathbf{Q}$, this allows the agent attention to model the global information with lower computational cost. Compared to traditional SA with a computational complexity of $O(N^2d)$ and LA with a computational complexity of $O(Nd^2)$, agent self-attention has a computational complexity of $O(Nnd)$. Overall, the computational complexity of agent attention is equivalent to generalized LA, which has the advantages of low computational complexity and strong feature representation ability.

The specific calculation steps of agent attention are as follows. First, the input tensor $\mathbf{X} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ applies 1×1 pointwise convolution to aggregate per-pixel channel information. Then, 3×3 depthwise convolutions encode spatial information, generating $\mathbf{Q}, \mathbf{A}, \mathbf{K}, \mathbf{V}$. The semantic agent $\mathbf{A}$ is obtained through AdaptiveAvgPool. After the reshape operation, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times d}$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$ are obtained as follows:

$$\begin{aligned}\mathbf{Q} &= \mathbf{W}_1^Q\mathbf{W}_3^Q\mathbf{X}\\ \mathbf{K} &= \mathbf{W}_1^K\mathbf{W}_3^K\mathbf{X}\\ \mathbf{V} &= \mathbf{W}_1^V\mathbf{W}_3^V\mathbf{X}r\\ \mathbf{A} &= \text{AvgPool}\left(\mathbf{W}_1^A\mathbf{W}_3^A\mathbf{X}\right)\end{aligned} \tag{11}$$

where $n \ll N$ and $\mathbf{W}_1^{()}$ and $\mathbf{W}_3^{()}$ are 1×1 pointwise convolutions and 3×3 depthwise convolutions, respectively. After obtaining $\mathbf{Q}, \mathbf{A}, \mathbf{K}, \mathbf{V}$, instead of computing the similarity between $\mathbf{Q}$ and $\mathbf{K}$ as in traditional attention, the semantic agent $\mathbf{A}$ with a smaller number of tokens is used to aggregate $\mathbf{Q}$ and $\mathbf{K}$ to obtain $\hat{\mathbf{Q}}$ and $\hat{\mathbf{K}}^{\text{T}}$. Subsequently, information is aggregated from $\hat{\mathbf{K}}^{\text{T}}$ and $\mathbf{V}$ and presented to $\hat{\mathbf{Q}}$. This design can guarantee global information modeling at a very low computational cost

$$\hat{\mathbf{Q}} = \sigma\left(\mathbf{Q}\mathbf{A}^T\right) \triangleq \phi_q(\mathbf{Q}) \tag{12}$$

$$\hat{\mathbf{K}}^T = \sigma\left(\mathbf{A}\mathbf{K}^T\right) \triangleq \phi_k(\mathbf{K}) \tag{13}$$

$$\text{ASA} = \hat{\mathbf{Q}}\hat{\mathbf{K}}^T\mathbf{V} \triangleq \phi_q(\mathbf{Q})(\phi_k(\mathbf{K})\mathbf{V}) \tag{14}$$

where $\sigma()$ denotes softmax, $\triangleq$ denotes the equivalence written as $\hat{\mathbf{Q}} \in \mathbb{R}^{N\times d}$ and $\hat{\mathbf{K}} \in \mathbb{R}^{n\times N}$, with a total computational complexity of $O(Nnd) \ll O(N^2d)$. It demonstrates that agent attention effectively combines traditional SA and LA, providing a strong ability for feature representation at a low computational cost.

In Fig. 6(c), GFFN splits operations into two branches: one employs the GELU activation function for gating and

enhancing important features, while the other directly processes local context information. The gated mechanism effectively suppresses irrelevant or redundant features and aggregates information from different levels, allowing each level to focus on complementing the details of other levels. The GFFN is expressed as follows:

$$\mathrm{GFFN} = \mathbf{W}_1(\varphi(\mathbf{W}_1\mathbf{W}_3(\mathrm{LN}(\mathbf{F}_{\mathrm{ASA}}))) \odot \mathbf{W}_1\mathbf{W}_3(\mathrm{LN}(\mathbf{F}_{\mathrm{ASA}}))) \tag{15}$$

where φ denotes GELU, while $\mathbf{W}_1$ and $\mathbf{W}_1$ represent the 1×1 pointwise convolution kernel and the 3×3 depthwise convolution kernel, respectively.

D. Loss Function

RSICD is a pixel-level binary classification task. However, the ratio of changed regions is much less than that of the unchanged ones in most cases, which results in a class imbalance problem. To alleviate this issue, we adopt a hybrid loss, including a binary cross-entropy (BCE) loss L_{bce} and a dice loss L_{dice}. The hybrid loss can be formulated as follows:

$$L_{\mathrm{bce}} = -y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \tag{16}$$

$$L_{\mathrm{dice}} = 1 - (2y_i p_i)/(y_i + p_i) \tag{17}$$

$$L_{\mathrm{cd}} = L_{\mathrm{bce}} + L_{\mathrm{dice}} \tag{18}$$

where y_i denotes the ground truth of the ith pixel and p_i denotes the change prediction of the ith pixel.

IV. Experiment and Results

A. Datasets

To evaluate the proposed method, experiments are conducted on three publicly available large RSICD datasets, including SYSU-CD [11], WHU-CD [60], and LEVIR-CD+ [10], following mainstream classical dataset partitioning methods to ensure fair experimental comparisons.

SYSU-CD is a public large-scale RSICD dataset. It contains 20 000 pairs of RS images of size 256×256 with 0.5-m resolution. In this article, we follow its default dataset split for experiments, 12 000/4000/4000 pairs of image patches are used for training/validation/testing. It is worth noting that the SYSU-CD dataset covers various target types other than buildings, such as vessels, roads, and vegetation.

WHU-CD is a public RSICD dataset focused on buildings. It contains one pair of RS images with a size of $32\,507 \times 15\,354$ and a resolution of 0.2 m. We follow the processing of the mainstream approach to crop the original image into small patches of size 256×256 without overlap and randomly split them into 6096/762/762 for training/validation/testing.

LEVIR-CD+ is a public large-scale RSICD dataset focused on buildings. It includes 637 and 985 pairs of RS images, each with a size of 1024×1024 pixels and a resolution of 0.5 m. In this article, we follow the data split of the original dataset and only crop the images into small patches of size 256×256. 10 192/5568 pairs of image patches are used for training/testing.

B. Implementation Details and Evaluation Metrics

The proposed M2M-LINet is implemented by PyTorch and trained using NVIDIA GeForce RTX 3090 GPU for 200 epochs. During the training process, we used the AdamW optimizer and set the momentum to 0.99. The weight decay is set to 0.0005 and the initial learning rate is 0. 0001. The λ_1 and λ_2 parameters in the LCB are adaptively learning and not manually set. We mainly used four evaluation metrics for the comprehensive evaluation of the proposed method, including precision (Pre), recall (Rec), $F1$-score, and DIP. They are formulated as follows:

$$\mathrm{Pre} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{19}$$

$$\mathrm{Rec} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{20}$$

$$F1 = \frac{2 \times \mathrm{Pre} \times \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}} \tag{21}$$

$$\mathrm{DIP} = 1 - \sqrt{\frac{(1 - \mathrm{Pre})^2 + (1 - \mathrm{Rec})^2}{2}} \tag{22}$$

where TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively. Among them, the $F1$ and DIP metrics are integrated with Prec and Rec, which are the main metrics.

C. Comparison With State-of-the-Art Methods

To verify the superiority of the proposed method, the M2M-LINet is compared with several state-of-the-art RSICD methods, which can be roughly categorized into three groups. First, CNN-based methods: FC-EF [5], FC-Diff [5], FC-Conc [5], FCN-PP [1], STANet [10], IFNet [6], FDCNN [9], SNUNet [8], and DSAMNet [11]. Second, Transformer-based methods: BIT [20]. Finally, hybrid architectures based on the CNN and the Transformer: ICIFNet [28], WNet [29], TCD [30], and EATDer [31].

1) Quantitative Evaluation: To verify the effectiveness of the proposed method, we conducted experiments on three large public datasets. As shown in Table I, our proposed M2M-LINet significantly outperforms other competitive methods based on the CNN or the Transformer on the three datasets. The experimental results on the SYSU-CD dataset show that the $F1$ score of the proposed method is 5.85% higher than the best CNN-based method DSAMNet and 3.15% higher than the best hybrid architecture-based method EATDer. On the WHU-CD dataset, the $F1$ score of the proposed method is 7.91% higher than the best CNN-based method DSAMNet, and 2.77% higher than the best hybrid architecture-based method WNet. On the LEVIR-CD+ dataset, the $F1$ score of the proposed method is 2.72% higher than the best CNN-based method IFN and 1.36% higher than the best hybrid architecture-based method ICIF-Net.

2) Qualitative Evaluation: The visual analysis of competitive methods on the three datasets is shown in Fig. 7. The proposed method shows significant advantages in the following aspects.

TABLE I

QUANTITATIVE COMPARISONS IN TERMS OF PRE, REC, $F1$, AND DIP ON THREE RSICD DATASETS. $F1$ AND DIP ARE COMPREHENSIVE METRICS; HIGHER $F1$ AND DIP VALUES INDICATE BETTER MODEL PERFORMANCE. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	Publish	SYSU-CD				WHU-CD				LEVIR-CD+			
		Pre(%)	Rec(%)	F1(%)	DIP(%)	Pre(%)	Rec(%)	F1(%)	DIP(%)	Pre(%)	Rec(%)	F1(%)	DIP(%)
FC-EF [5]	ICIP 2018	72.28	72.85	72.57	72.56	78.79	68.15	73.08	72.94	70.60	73.64	72.09	72.08
FC-Diff [5]	ICIP 2018	72.69	79.46	75.93	75.84	80.44	54.59	65.04	65.04	79.94	72.55	76.06	75.96
FC-Conc [5]	ICIP 2018	82.17	72.06	76.78	76.56	75.89	69.30	72.61	72.40	78.07	68.22	72.82	72.70
FCNPP [1]	GRSL2019	74.60	78.91	76.70	76.66	79.94	89.76	84.57	84.07	72.49	74.44	73.45	73.45
STANet [10]	RS 2020	70.76	85.33	77.37	76.87	79.37	85.50	82.32	82.17	69.74	83.92	76.17	75.77
IFNet [6]	ISPRS 2020	79.59	73.58	76.47	76.39	78.00	70.81	74.23	74.15	83.77	80.32	82.29	81.96
FDCNN [9]	TGRS 2020	78.43	74.54	76.44	76.40	85.71	84.42	85.06	85.05	73.84	78.96	76.31	76.26
SNUNet [8]	GRSL 2022	78.68	76.37	77.51	77.50	85.58	80.99	83.22	83.13	80.52	79.43	80.82	79.96
DSAMNet [11]	TGRS 2022	74.81	81.86	78.18	78.05	84.43	87.86	86.11	86.04	80.73	76.52	78.57	78.52
BIT [20]	TGRS 2022	78.73	75.68	77.17	77.15	88.70	86.26	87.46	87.42	82.74	82.85	82.80	82.80
ICIF-Net [28]	TGRS 2022	83.37	78.51	80.74	80.79	92.98	85.56	88.32	88.65	87.79	80.88	**83.65**	**83.96**
WNet [29]	TGRS 2023	81.71	79.58	80.64	80.62	92.37	90.15	**91.25**	**91.19**	80.18	78.68	79.43	79.42
TCD [30]	TGRS 2023	75.25	84.23	79.49	79.25	87.39	90.85	89.09	88.98	79.96	84.85	82.34	82.24
EATDer [31]	TGRS 2024	79.82	81.96	**80.88**	**80.86**	88.74	91.32	90.01	89.95	78.13	80.97	79.52	79.50
M2M-LINet	/	82.95	85.14	**84.03**	**84.01**	92.89	95.17	**94.02**	**93.92**	82.77	87.37	**85.01**	**84.89**

a) Advantage in location accuracy: The proposed M2M-LINet performs particularly well in complex scenes. In the SYSU-CD dataset, where vegetation and road changes are complex and irregular, M2M-LINet effectively locates the real change regions. In the WHU-CD dataset, due to the similarity between buildings and background features, it is difficult for other methods to accurately locate the changes of buildings, which leads to missed detection. In contrast, M2M-LINet can accurately detect these changes and effectively suppress background interference. Similarly, in the LEVIR-CD+ datasets, M2M-LINet significantly improves target localization accuracy, accurately identifying changed areas even in complex backgrounds.

b) Advantage in distinguishing pseudo-changes: In the SYSU-CD dataset, many methods produce false detections due to changes in shooting angles and irrelevant interference around ships (marked by yellow boxes). However, M2M-LINet delivers the best detection results. In the WHU-CD dataset, irrelevant road and vehicle changes (yellow boxes) are successfully avoided by M2M-LINet. In the LEVIR-CD+ dataset, M2M-LINet successfully avoids irrelevant changes such as road and pool changes (yellow boxes).

c) Advantage in content integrity: M2M-LINet performs well in identifying the integrity of change areas and can adapt to various types of change scenes. In the SYSU-CD dataset, which contains irregular change scenes, M2M-LINet can entirely detect the change areas. In the WHU-CD dataset, M2M-LINet can capture the overall changes when facing large regular building changes, while other methods fail to fully recognize these complex changes. In the LEVIR-CD+ dataset, M2M-LINet also shows excellent detection capability for small target changes and provides the most complete detection results.

These experimental results show that the proposed M2M-LINet method significantly enhances performance in RSICD. The improvement is attributed to two key aspects. First, the effectiveness of the three-stage architecture. M2M-LINet adopts a multistage strategy from coarse localization to fine-detail aggregation and gradually integrates multilevel temporal difference features. It provides more fine-grained change maps for the RSICD task. Second, the auxiliary decoding BAEM module introduces boundary information for assistance, which ensures better integrity of the change information.

D. Ablation Analysis

To validate the effectiveness of each proposed module, we conducted ablation experiments of components on the three datasets. As shown in Table II, "w/o" denotes the removal module, "w" denotes the usage module, and the N01 denotes the M2M-LINet method proposed in this article. The detailed analysis of each key module is presented as follows.

1) Effectiveness of CALM: To verify the effectiveness, we first removed the CALM module (N02 in Table II), which resulted in a significant performance decrease on the three datasets. Second, the CALM was replaced with ADConv convolution (N03 in Table II). From the quantitative results in Table II, it can be seen that either removing or replacing

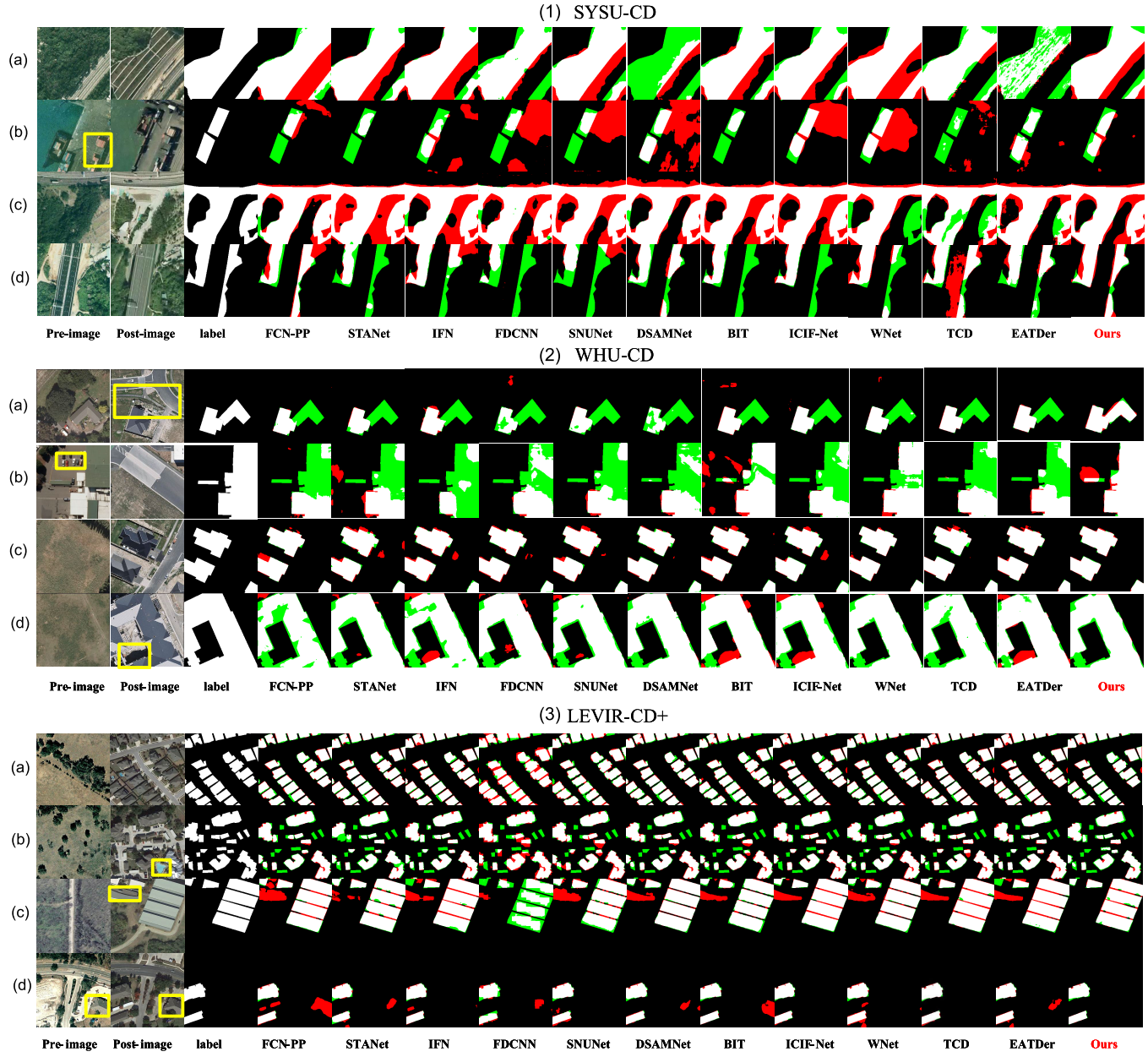

Fig. 7. Visualization comparison of results on three datasets. Each dataset displays four different sets of visualization results, labeled as (a)–(d). In each set of images, yellow boxes highlight irrelevant changes in the data that are considered interference factors. Interference factors may include: unrelated road changes, vehicle changes, or building shadow changes. The rendered colors represent true positives (white), false positives (red), true negatives (black), and false negatives (green). The less red and green, the more accurate the detection.

the module leads to a performance decrease. It demonstrates that aggregated features are essential for the effective location of the changed region in the coarse localization stage.

2) Effectiveness of LCB: After removing LCB (N04 in Table II), high-frequency enhancement (N05 in Table II), and low-frequency enhancement (N06 in Table II), the detection performance of the model significantly decreased. In addition, when the LCB was replaced by the convolutional layer (N07 in Table II), the detection accuracy $F1$ decreased by about 0.82%/1.07%/1.7% on the three datasets, respectively. This shows that LCB plays an important role in enhancing the image information and improving the detection performance.

3) Effectiveness of LTB: After removing LTB (N08 in Table II), the detection performance on three datasets significantly decreased. After removing MASA (N09 in Table II), the $F1$ scores on the three datasets decreased by 0.58%, 1.3%, and 1.33%, respectively. Additionally, we also replaced the GFFN in LTB with the traditional FFN (N10 in Table II), and the results also showed a performance decline. Similarly, we replaced the MASA in LTB with LA (N11 in Table II). Although the computational complexity and parameter number are approximate, the detection accuracy $F1$ is decreased by 0.43%/0.82%/0.44% on the three datasets, respectively. These results demonstrate the effectiveness and advantages of

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

Net		Efficiency		SYSU-CD			WHU-CD			LEVIR-CD+		
		Params	FLOPS	Pre (%)	Rec (%)	F1(%)	Pre (%)	Rec (%)	F1(%)	Pre (%)	Rec (%)	F1(%)
N01	M2M-LINet	6.554M	3.351G	82.95	85.14	84.03	92.89	95.17	94.02	82.77	87.37	85.01
(a) **Stage1**: Coarse Aggregation Localization Module (CALM)												
N02	w/o CALM	6.458M	3.134G	81.11	84.52	82.78	89.21	95.33	92.17	79.25	87.11	82.99
N03	w ADConv	6.511M	3.252G	83.04	83.44	83.24	89.01	95.31	92.05	81.85	85.10	83.45
(b) **Stage2**: Lightweight Convolutional Block (LCB)												
N04	w/o LCB	5.929M	3.115G	80.59	85.25	82.86	88.67	94.03	91.27	83.43	82.94	83.19
N05	w/o high	6.272M	3.240G	81.09	86.31	83.62	92.14	94.39	93.25	84.00	83.22	83.61
N06	w/o low	6.272M	3.296G	81.16	85.66	83.35	89.80	95.73	92.67	83.72	83.16	83.44
N07	w CNN	5.976M	3.168G	84.28	82.16	83.21	89.78	96.35	92.95	84.21	82.44	83.31
(c) **Stage2**: Lightweight Transformer Block (LTB)												
N08	w/o LTB	4.832M	2.290G	79.25	86.78	82.84	87.90	95.87	91.71	82.18	83.15	82.66
N09	w/o MASA	5.632M	2.680G	83.50	83.40	83.45	90.79	94.73	92.72	84.38	82.99	83.68
N10	w FFN	6.504M	3.262G	81.43	86.02	83.66	92.49	94.51	93.49	82.32	86.26	84.24
N11	w LA	6.491M	3.138G	83.40	83.81	83.60	92.27	94.14	93.20	84.43	84.71	84.57
(d) **Stage3**: Boundary-Aware Enhancement Module (BAEM)												
N12	w/o BAEM	6.403M	3.187G	82.98	83.80	83.38	89.70	96.27	92.87	81.84	85.85	83.79
N13	w/o Boundary	6.542M	3.367G	81.77	86.01	83.52	90.70	95.85	93.20	83.45	85.31	84.37
(e) Loss Function												
N14	w BCE	6.554M	3.351G	82.89	84.28	83.58	91.46	95.69	93.53	84.72	84.08	84.40
N15	w Dice	6.554M	3.351G	81.83	85.83	83.77	91.13	96.35	93.67	83.32	86.06	84.67

LTB in extracting global information and keeping the model lightweight.

4) Effectiveness of BAEM: After removing BAEM (N12 in Table II), the detection accuracy $F1$ is decreased by 0.75%/1.15%/1.04% on the three datasets, respectively. Additionally, after eliminating boundary information (N13 in Table II), the detection accuracy $F1$ is decreased by 0.51%/0.82%/0.64% on the three datasets, respectively. These results show that BAEM plays a critical role in the auxiliary decoding process, especially in the utilization of boundary information.

5) Discussions of the Loss Function: To verify the hybrid loss effectiveness, we conducted experiments to validate the methods using only BCE loss (N14 in Table II) or only Dice loss (N15 in Table II). The results show that the method using only the Dice loss function is better than the method using only the BCE loss function. This demonstrates that the hybrid loss has a more significant advantage in improving the performance of the model.

E. Discussion of Architecture and Model Efficiency

1) Architecture Effectiveness: To evaluate the effectiveness of the proposed architecture, we conducted a more detailed analysis and validation. *(1) Three-stage architecture*. First, the effectiveness of the three-stage CD framework is verified by Table II(a)–(d). Removing the coarse location stage to maintain consistency with the two-stage method of the traditional encoder–decoder structure, the experimental results show a significant decrease in model performance. It demonstrates the importance of the coarse location stage in the overall architecture. In addition, the visualized results in Fig. 8 further show how the three-stage architecture simulates the "global-first" processing approach of biological vision, gradually refining features from macro to micro. In the initial stage, the model first locates the change areas on a large scale, then gradually refines and enhances the features, visually demonstrating how the architecture captures and processes change information at each stage. *(2) Interleaved architecture*. In quantitative experiments, our method M2M-LINet is compared with mainstream serial design (e.g., BIT [20] in the comparative method) and parallel design (e.g., ICIFNet [28] and WNet [29]), which validate the effectiveness of the interleaved architecture. To further validate the interleaved architecture, we also performed serial and parallel designs for the components LCB and LTB proposed in this article and compared them with the experiments. The experimental results in Fig. 9 clearly show that the interleaved architecture is not only superior in performance but also has less

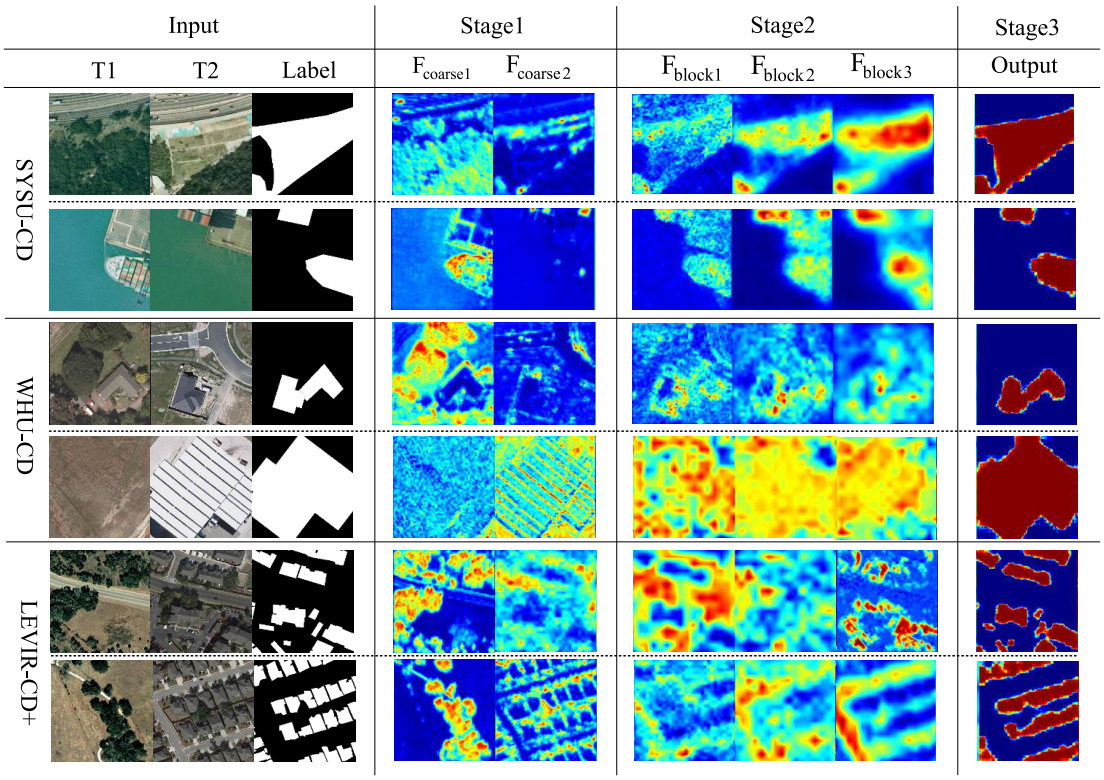

Fig. 8. Visual heat map of the three stages.

TABLE III

COMPARISON RESULTS OF COMPUTATIONAL EFFICIENCY. WE REPORT PARAMETERS (PARAMS.) AND FLOATING-POINT OPERATIONS (FLOPS), AS WELL AS THE $F1$ ON THREE RSICD TEST SETS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	F1(%)			FLOPs(G)	Params(M)
	SYSU	WHU	LEVIR+		
FCNPP	76.70	84.57	73.45	34.65	28.13
STANet	77.37	82.32	76.17	**6.58**	16.93
IFNet	76.47	74.23	82.29	41.18	50.71
FDCNN	76.44	85.06	76.31	32.40	13.71
SNUNet	77.51	83.22	80.82	33.04	12.03
DSAMNet	78.18	86.11	78.57	75.29	16.95
BIT	77.17	87.46	82.80	8.44	6.93
ICIF-Net	80.74	88.32	**83.65**	25.36	23.82
WNet	80.64	**91.25**	79.43	19.20	43.07
TCD	79.49	89.09	82.34	56.64	11.13
EATDer	**80.88**	90.01	79.52	23.46	**6.60**
M2M-LINet	**84.03**	**94.02**	**85.01**	**3.351G**	**6.554M**

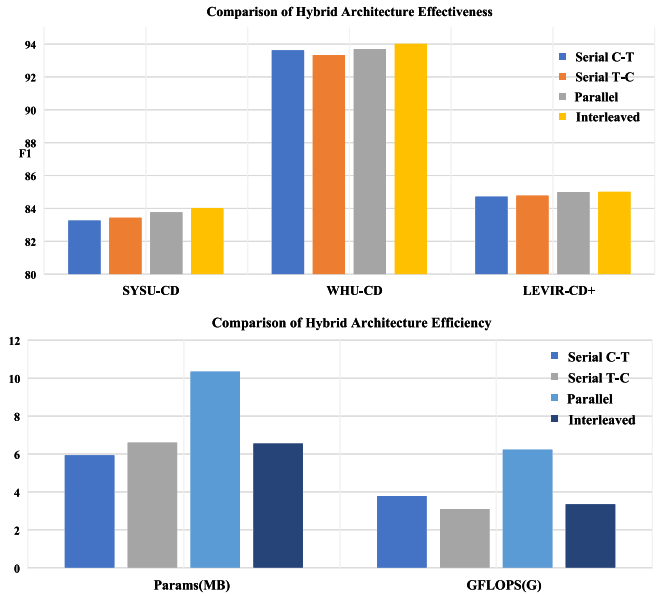

Fig. 9. Performance and efficiency comparison of different hybrid architectures. Different hybrid architectures are designed for the proposed LCB and LTB components, where Serial C-T represents LCB followed by LTB, Serial T-C represents LTB followed by LCB, Parallel represents dual branches of LCB and LTB, and Interleaved is the M2M-LINet proposed in this article. These four architectures correspond to the four hybrid architectures shown in Fig. 1.

computational resource consumption compared to other hybrid architectures.

2) Model Efficiency: We analyzed and compared the comparison methods from the perspectives of floating-point operations (FLOPs), number of parameters (Params), and $F1$ scores. The detailed results are shown in Table III, and the effective comparison is also given in Fig. 10. Since the methods [5] use fewer layers of feature extraction and their detection accuracy is low, we did not compare them. The results lead to the following conclusions: the proposed M2M-LINet method achieves optimal accuracy while

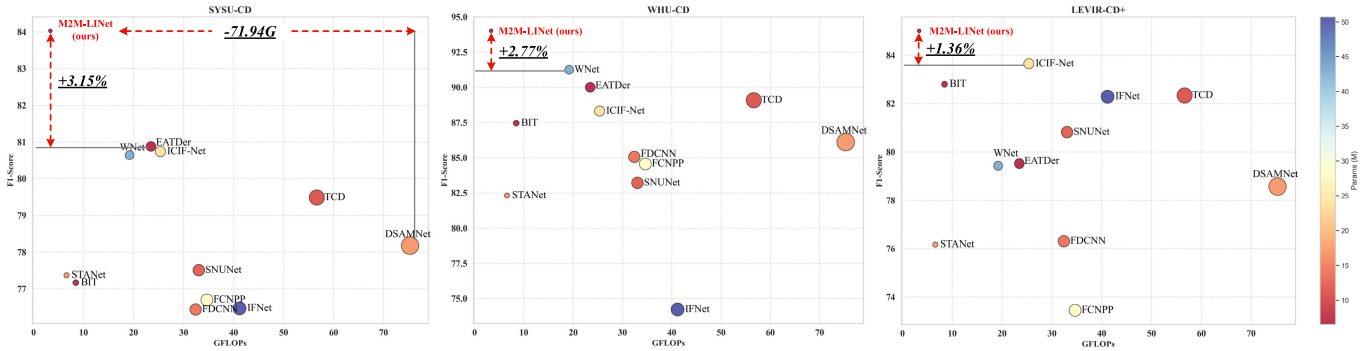

Fig. 10. The model complexity comparison of different methods in terms of parameters (memory cost), FLOPs (computational cost), and $F1$-score on three datasets. The proposed M2M-LINet method exhibits higher accuracy, lower parameters, and lower computational cost.

exhibiting the lowest computational complexity with only 3.351G FLOPs and the smallest parameter size of 6.554 MB compared to other mainstream change detection methods. Compared to EATDer [31], M2M-LINet has similar parameter numbers but only 14% of its computational complexity. These results demonstrate the effectiveness and superiority of the proposed method on model efficiency.

V. Conclusion

In this article, we have proposed a lightweight M2M-LINet method for RSICD tasks. The proposed method addresses the main problems in the current RSICD by designing a new three-stage change detection framework and a lightweight interleaved architecture. Specifically, the three-stage framework gradually improves the change information from coarse to fine. First, the coarse location stage helps to obtain the rough range and location of the change. Second, the fine detail focusing stage further extracts the coarse-grained information in the first stage to obtain finer details of the changes. Finally, the decoding prediction stage generates the final change result. The lightweight interleaved architecture enables effective local–global information interaction through LCB and LTB components and maintains the lightness of the method. The experimental results on three change detection public datasets show that our method outperforms other competitive methods in terms of change detection accuracy, number of parameters, and FLOPs.

References

[1] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.

[2] K. Chen et al., "RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701117.

[3] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023, Art. no. 5622018.

[4] J. Pan, Y. Bai, Q. Shu, Z. Zhang, J. Hu, and M. Wang, "M-swin: Transformer-based multiscale feature fusion change detection network within cropland for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024, Art. no. 4702716.

[5] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[6] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[7] Y. Zhang, L. Fu, Y. Li, and Y. Zhang, "HDFNet: Hierarchical dynamic fusion network for change detection in optical aerial images," *Remote Sens.*, vol. 13, no. 8, p. 1440, Apr. 2021.

[8] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[9] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.

[10] 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, p. 1662, 2020.

[11] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, Art. no. 5604816.

[12] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention Siamese network for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406216.

[13] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

[14] T. Lei et al., "Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 4507013.

[15] Z. Li et al., "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, Art. no. 5602812.

[16] T. Lei et al., "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-14, 2023, Art no. 4402114.

[17] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[18] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[19] F. Shamshad et al., "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802.

[20] 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, Art. no. 5607514.

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

[22] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022, Art. no. 5622519.

[23] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[24] T. Yan, Z. Wan, P. Zhang, G. Cheng, and H. Lu, "TransY-Net: Learning fully transformer networks for change detection of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, Art. no. 4410012.

[25] X. Xu, Z. Yang, and J. Li, "AMCA: Attention-guided multiscale context aggregation network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–19, 2023, Art. no. 5908619.

[26] R. Song, W. Ni, W. Cheng, and X. Wang, "CSANet: Cross-temporal interaction symmetric attention network for hyperspectral image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[27] X. Zhang, S. Cheng, L. Wang, and H. Li, "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, Art. no. 2000415.

[28] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 4410213.

[29] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "WNet: W-shaped hierarchical network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023, Art. no. 5615814.

[30] D. Xue et al., "Triple change detection network via joint multifrequency and full-scale swin-transformer for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4408415.

[31] J. Ma, J. Duan, X. Tang, X. Zhang, and L. Jiao, "EATDer: Edge-assisted adaptive transformer detector for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024, Art. no. 5602015.

[32] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognit. Psychol.*, vol. 9, no. 3, pp. 353–383, Jul. 1977.

[33] N. Khabou, I. B. Rodriguez, G. Gharbi, and M. Jmaiel, "A threshold based context change detection in pervasive environments: Application to a smart campus," *Proc. Comput. Sci.*, vol. 32, pp. 461–468, Jan. 2014.

[34] G. Yang, H.-C. Li, W.-Y. Wang, W. Yang, and W. J. Emery, "Unsupervised change detection based on a unified framework for weighted collaborative representation with RDDL and fuzzy clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8890–8903, Nov. 2019.

[35] 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.

[36] Z. Li, W. Shi, H. Zhang, and M. Hao, "Change detection based on Gabor wavelet features for very high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 783–787, May 2017.

[37] P. Lu, A. Stumpf, N. Kerle, and N. Casagli, "Object-oriented change detection for landslide rapid mapping," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 4, pp. 701–705, Jul. 2011.

[38] X. Zhang, P. Xiao, X. Feng, and M. Yuan, "Separate segmentation of multi-temporal high-resolution remote sensing images for object-based change detection in urban area," *Remote Sens. Environ.*, vol. 201, pp. 243–255, Nov. 2017.

[39] Y. Shangguan, J. Li, Z. Chen, L. Ren, and Z. Hua, "Multiscale attention fusion graph network for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024, Art. no. 4402618.

[40] K. Song and J. Jiang, "AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4816–4831, 2021.

[41] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art. no. 4412712.

[42] H. Zheng et al., "HFA-net: High frequency attention Siamese network for building change detection in VHR remote sensing images," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108717.

[43] Z. Li et al., "STADE-CDNet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024, Art. no. 5611617.

[44] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, "Multiscale diff-changed feature fusion network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, Art. no. 5502713.

[45] Z. Ying et al., "DGMA2-Net: A difference-guided multiscale aggregation attention network for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024, Art. no. 5619716.

[46] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[47] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[48] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023.

[49] 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.

[50] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 599–609, Aug. 2023.

[51] T. Lei et al., "Lightweight structure-aware transformer network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[52] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[53] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.

[54] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[55] L. Yang, R. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 11863–11874.

[56] A. Hatamizadeh et al., "FasterViT: Fast vision transformers with hierarchical attention," 2023, *arXiv:2306.06189*.

[57] W. Zhang et al., "TopFormer: Token pyramid transformer for mobile semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12073–12083.

[58] Y. Chen et al., "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5260–5269.

[59] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "SwiftFormer: Efficient additive attention for transformer-based real-time mobile vision applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17379–17390.

[60] 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.

Yetong Xu received the B.S. degree in engineering from Shaanxi University of Science and Technology, Xi'an, China, in 2022, where she is currently pursuing the M.S. degree in engineering with the School of Electronic Information and Artificial Intelligence.

Her research interests include image processing and pattern recognition.

Tao Lei (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2011.

From 2012 to 2014, he was a Post-Doctoral Research Fellow with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an. From 2015 to 2016, he was a Visiting Scholar with the Quantum Computation and Intelligent Systems Group, University of Technology Sydney, Australia. From July 2017 to October 2017, he was a Research Fellow with the College of Electronic and Computer Engineering, Brunel University of London, U.K. From November 2023 to April 2024, he was a Research Fellow with the School of Information Science and Technology, Aichi University, Japan. He has authored and co-authored more than 100 research articles. He is currently a Professor with Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an. His research interests include image processing, pattern recognition, and machine learning.

Hailong Ning received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, P. R. China, in 2021.

He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, P. R. China. His main research interests include pattern recognition, machine learning, computer vision, and multi-modal learning.

Shaoxiong Lin received the B.S. degree in engineering from Shaanxi University of Science and Technology, Xi'an, China, in 2023, where he is currently pursuing the M.S. degree with the School of Electronic Information and Artificial Intelligence.

His research interests include image processing and pattern recognition.

Tongfei Liu (Member, IEEE) received the master's degree from the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China, in 2020, and the Ph.D. degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2023.

He is currently a Lecturer with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an. His research interests include deep learning, spatial–spectral feature extraction, pattern recognition, building extraction, ang land cover change detection and classification, through VHR remote sensing images (including satellite and aerial images).

Maoguo Gong (Fellow, IEEE) received the B.S. degree (Hons.) in electronic engineering and the Ph.D. degree in electronic science and technology from Xidian University, Xi'an, China, in 2003 and 2009, respectively.

Since 2006, he has been a Teacher with Xidian University, where he was promoted as an Associate Professor and a Full Professor, in 2008 and 2010, respectively, with exceptive admission. His research interests are in the areas of computational intelligence with applications to optimization, learning, data mining, and image understanding.

Dr. Gong received the Prestigious National Program for the Support of Top-Notch Young Professionals from the Central Organization Department of China, the Excellent Young Scientist Foundation from the National Natural Science Foundation of China, and the New Century Excellent Talent in University from the Ministry of Education of China. He is also an Associate Editor of IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Asoke K. Nandi (Life Fellow, IEEE) received the Ph.D. degree in physics from the University of Cambridge (Trinity College), Cambridge, U.K., in 1979.

He held academic positions in several universities, including Oxford, U.K., Imperial College London, U.K., Strathclyde, Glasgow, U.K., and Liverpool, U.K., as well as Finland Distinguished Professorship. In 2013, he moved to Brunel University of London, Uxbridge, U.K. In 1983, he co-discovered the three fundamental particles known as W^+, W^- and Z^0, providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics in 1984 awarded the prize to two of his team leaders for their decisive contributions. He made fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "Big Data." He has authored over 650 technical publications, including 310 journal articles as well as six books, entitled *Image Segmentation: Principles, Techniques, and Applications* (Wiley, 2022), *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines* (Wiley, 2020), *Automatic Modulation Classification: Principles, Algorithms and Applications* (Wiley, 2015), *Integrative Cluster Analysis in Bioinformatics* (Wiley, 2015), *Blind Estimation Using Higher-Order Statistics* (Springer, 1999), and *Automatic Modulation Recognition of Communications Signals* (Springer, 1996). The H-index of his publications is 91 (Google Scholar) and the ERDOS number is 2. His research interests lie in signal processing and machine learning, with applications to machine health monitoring, functional magnetic resonance data, gene expression data, communications, and biomedical data.

Dr. Nandi is a Fellow of the Royal Academy of Engineering and a Fellow of six other institutions including the IEEE. In 2023, he was honored by the Academia Europaea and the Academia Scientiarum et Artium Europaea. He has received many awards, including the IEEE Heinrich Hertz Award, in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers in 1999, and the Mountbatten Premium of the Institution of Electrical Engineers in 1998. He was an IEEE Distinguished Lecturer (EMBS, from 2018 to 2019).