Mask-Guided Local–Global Attentive Network for Change Detection in Remote Sensing Images

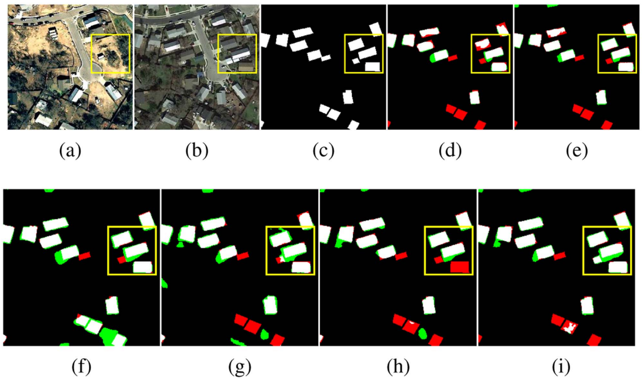
Fengchao Xiong, *Member, IEEE*, Tianhan Li, Jingzhou Chen, Jun Zhou, *Senior Member, IEEE*, and Yuntao Qian, *Senior Member, IEEE*

***Abstract*—Change detection in remote sensing images is a challenging task due to object appearance diversity and the interference of complex backgrounds. Self-attention- and spatial-attention-based solutions face limitations, such as high memory consumption and an inadequate ability to capture long-range relations, leading to imprecise contextual information and restricted performance. To address these challenges, this article introduces a novel mask-guided local–global attentive network (MLA-Net). The MLA-Net incorporates a memory-efficient local–global attention module that leverages the benefits of both self-attention and spatial attention to accurately capture the local–global context. Through simultaneous exploitation of context within inter- and intrapatches and information refinement, the feature representation capability is significantly enhanced. In addition, we introduce a change mask to refine feature differences and eliminate interference from irrelevant changes caused by complex backgrounds. Accordingly, a mask loss is defined to guide the generation of the mask. Extensive experiments on the LEVIR-CD, WHU-CD, and CLCD datasets show that our MLA-Net performs better than state-of-the-art methods.**

***Index Terms*—Attention mechanism, change detection (CD), change mask, convolutional neural network (CNN), remote sensing image.**

I. Introduction

WITH global climate change, people around the world are more concerned about the earth than ever before [1]. Remote sensing (RS) data provide an unbiased, uninterrupted, and unbounded view of human activities and natural processes. By comparing pairs of images taken at different moments, RS change detection (CD) assigns pixel-level binary labels indicating a change or no change and plays an important role in earth observation. It has driven many practical applications, including damage assessment, urban planning, and environmental monitoring [2], [3], [4], [5].


(a) (b) (c) (d) (e)

(f) (g) (h) (i)

Fig. 1. Visual sample results of different methods. (a) T1 images. (b) T2 images. (c) Ground truth. (d) BIT. (e) ChangeFormer. (f) EGRCNN. (g) STANet. (h) ICIF-Net. (i) Ours. (Red: false negative; Green: false positive.)

Although CD performance has been improved by deep learning (DL), it still faces several challenges.

1) *High diversity of target appearance:* Differences in imaging and lighting conditions, along with seasonal variations, lead to objects with the same semantic meaning displaying varying colors and shapes at different times and spatial locations. For instance, as illustrated in Fig. 1, the same building in T1 and T2 exhibits different colors and brightness. Even within the same scene, the appearance of buildings may differ. Overcoming this challenge requires the extraction of highly discriminative contextual information to identify the changes of interest.
2) *Background complexity:* Interference from clouds, haze, and noise is likely to introduce pseudo changes. This necessitates the extraction of accurate feature differences to mitigate the impact of unwanted changes.

Based on the human cognitive system, various attention mechanisms, such as spatial attention, channel attention, and self-attention [6], [7], [8], [9], [10], [11], [12], [13], have been developed to enhance feature discriminative capability. Albeit the ability to calculate long-range interactions, self-attention is limited by high memory requirement and is often deployed on low-resolution features, making it unlikely to obtain accurate context information effectively. Because of long-distance imaging, RS objects usually have limited details, a factor partially

Manuscript received 7 December 2023; revised 19 December 2023; accepted 31 December 2023. Date of publication 5 January 2024; date of current version 22 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62002169 and Grant 62071421, in part by the State Key Laboratory of Geo-Information Engineering and the Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, Chinese Academy of Surveying and Mapping, and in part by the Fundamental Research Funds for the Central Universities under Grant 30923010213. *(Corresponding author: Jingzhou Chen.)*

Fengchao Xiong, Tianhan Li, and Jingzhou Chen are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: fcxiong@njust.edu.cn; litianh0@qq.com; jzchen@njust.edu.cn).

Jun Zhou is with the School of Information and Communication Technology, Griffith University, Nathan 4111, QLD, Australia (e-mail: jun.zhou@griffith.edu.au).

Yuntao Qian is with the College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: ytqian@zju.edu.cn).

The code and models will be publicly available at https://github.com/bearshng/mla-net for reproducible research.

Digital Object Identifier 10.1109/JSTARS.2024.3350044

overlooked by self-attention that always pays extensive attention to global information. Although feature maps can be recalibrated by spatial and channel attention to highlight the important parts, they struggle to enjoy the benefits of long-range contexts. To the best of our knowledge, it is still underexplored to effectively encode both local characteristics and global semantic interactions while suppressing background interference for CD.

There is comparatively less work on the accurate extraction of feature differences compared with the large amount of work on feature representation. Most work performs feature subtraction between bitemporal images. However, background noise, scale differences, etc., can mislead to pseudo changes and result in false alarms [14], [15]. One way is to use recurrent neural networks (RNNs) to capture spatial–temporal changes [16], [17]. However, the suitability of modeling bitemporal images using RNNs is questionable, as RNNs were originally designed for sequential data rather than bitemporal data. An alternative approach is to use spatialwise and channelwise attention to refine feature changes [15], [18], [19], [20]. For example, the MDESNet [21] applies the convolutions and sigmoid activation function to obtain a one-channel feature difference attention map, which is used to calibrate the concatenated features. In the absence of supervision, it is unclear whether the attention map can force the network to extract the accurate feature differences.

In view of the above problems, this article proposes a mask-guided local–global attentive network (MLA-Net) for CD from the following two aspects

1) *More discriminative contextual feature representation:* We design a memory-efficient local–global attention (LGA) module by combining the advantages of spatial attention and self-attention. Instead of pixelwise self-attention computation across the image, our LGA module builds the local self-attention within patches and global self-attention between patches to simultaneously capture the local and global contexts. Moreover, the two attentions are further blended, producing a weight to further filter out the irrelevant background information.
2) *More accurate feature differences extraction:* We introduce a change mask to refine the initial feature differences computed via feature subtraction. An additional mask loss is introduced to generate the mask more accurately. In this way, unreliable pseudo-feature differences are greatly suppressed. Thanks to the hybrid advantages of the above techniques, our method achieves state-of-the-art performance on the LEVIR-CD, WHU-CD, and CLCD datasets.

To summarize, the contributions of this article are threefold.

1) We introduce a memory-efficient LGA module for robust contextual information extraction. This module effectively leverages both local and global contexts while simultaneously filtering out irrelevant information. Its low memory cost makes it feasible for application in high-resolution feature maps, enabling better handling of the high diversity of target appearances and the complexity of backgrounds.
2) We develop a change mask and the corresponding mask loss to ensure the accurate extraction of feature differences. The change mask is produced in a supervised manner, guiding the precise generation of feature differences. Consequently, this approach significantly suppresses pseudo and unwanted changes, contributing to more accurate results.
3) Our MLA-Net has undergone extensive comparisons with state-of-the-art methods and achieved the F1 scores of 91.72%, 95.06%, and 79.88% on the LEVIR-CD, WHU-CD, and CLCD datasets, respectively. This showcases the superior performance of MLA-Net in CD tasks.

The rest of this article is organized as follows. Section II presents the related works. Our proposed approach is demonstrated in Section III. The experimental results are reported in Section IV. Discussion is presented in Section V. Finally, Section VI concludes this article.

II. Related Works

A. DL-Based RS CD Approaches

Traditional works use data transformations and image algebras, such as principal component analysis [22], change vector analysis [23], [24], and multivariate CD [25], for CD. The change map is obtained by performing threshold- and cluster-based methods on the difference images. By contrast, classification-based methods perform classification on bitemporal images and regard the pixels or regions belonging to different classes as changes [26]. However, these methods typically rely on handcrafted features with insufficient representation ability and are prone to atmospheric conditions, seasonal changes, and satellite sensors, greatly hindering CD accuracy.

Driven by the wide availability of RS images and high-level discriminative feature extraction ability, DL-based methods have substantially boosted the CD performance [27], [28], [29], [30]. Autoencoders [31], RNNs [32], convolutional neural networks (CNNs) [7], generative adversarial networks [33], and transformers [34] are widely used network architectures for discriminative hierarchical features extraction. CD can be regarded as the detection of semantic changes between temporal images. For this reason, fully convolutional networks [35], [36] and U-Net [37], [38] and their variants [39], [40] are modified for this task. Differentiated by the way to handle bitemporal images, early fusion and late fusion are typical approaches. Early fusion methods cascade the inputs for feature extraction, followed by classification. In contrast, late fusion uses a shared backbone network to extract features from bitemporal images individually and compare feature differences to detect changes. Compared to early fusion networks, late fusion networks can highlight the differences between images and obtain more competitive performance.

As aforementioned, RS objects have diverse appearances. To improve the separability of different objects, extensive efforts have been made to improve the feature representation capabilities of networks for CD [6], [41]. Dilated convolution can enlarge the receptive field and was adopted in [17] to enhance feature extraction ability. RS objects usually have irregular shapes and different scales, which requires multiscale feature extraction [12]. As such, Liu et al. [42] proposed a local–global pyramid network to extract more discriminant features for the buildings of different scales from global and local perspectives.

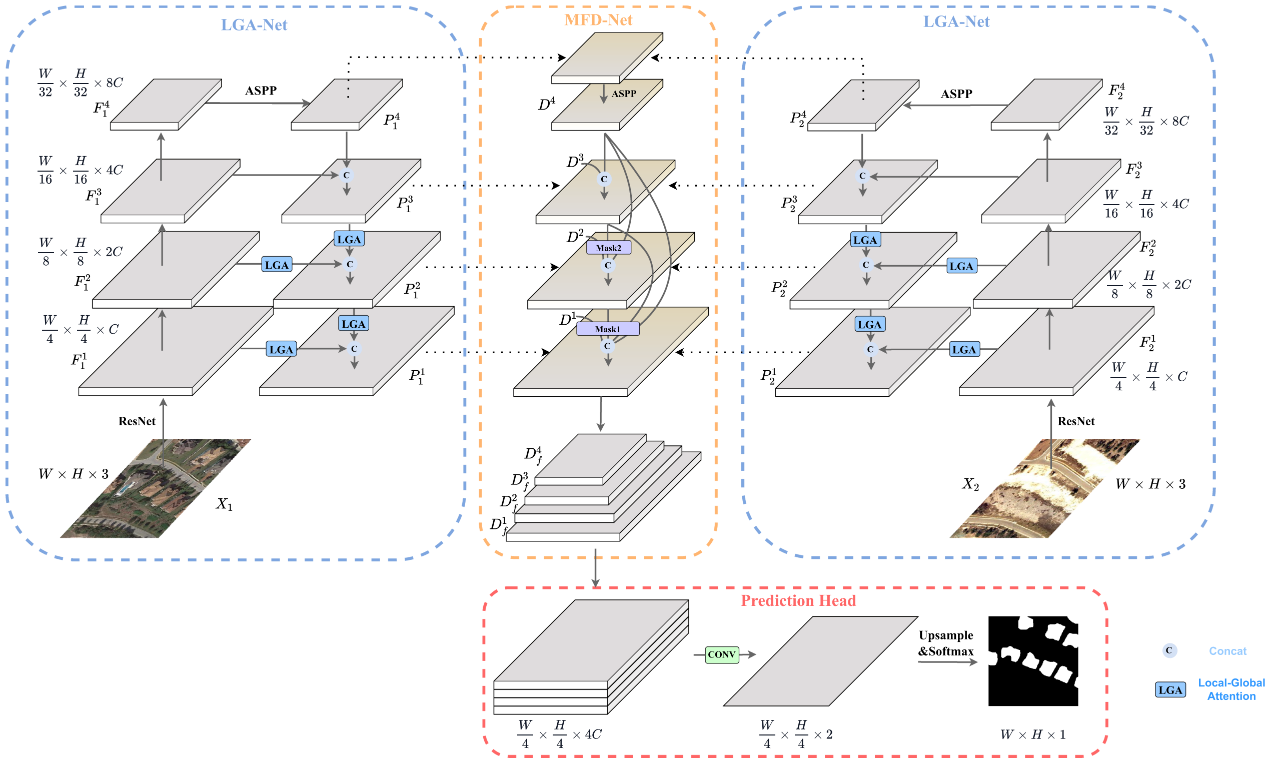


Fig. 2. Illustration of the proposed MLA-Net. It includes LGA-Net, MFD-Net, and prediction head to extract highly discriminative contextual information, produce accurate feature differences, and generate the change mask, respectively.

Zhang et al. [43] took advantage of Transformer in global information extraction and developed Transformer U-Net to enhance the feature extraction capability and achieved higher accuracy. Combining the advantages of CNNs and Transformer, Feng et al. [7] proposed an intrascale cross-interaction and interscale feature fusion network (ICIF-Net) to harvest the local–global features simultaneously. Following this line, we also focus on improving the feature extraction ability of the network but more efficiently and effectively.

B. Attention for RS Image CD

Driven by the success of the attention mechanism in many computer vision tasks, extensive efforts have been made in CD with attentions [44], [45]. On the one hand, attention allows the network to focus on the most critical features and locations, suppressing irrelevant features and locations related to background and noise [46], [47]. Channel and spatial attentions were used in [48] and [49] to refine the channelwise and spatialwise features, which greatly mitigates the effect of pseudo changes. On the other hand, attention is adopted to fuse the features at various levels, yielding semantically and contextually richer features. Zhang et al. [14] achieved multilevel context aggregation through a multilevel and cross-level attention fusion scheme. However, in spatial and channel attentions, the receptive field remains limited in capturing global context.

In contrast, by modeling pairwise long-range interactions between different image regions [50], [51], [52], self-attention can extract the global representation for the whole image and enhance the discriminative capability of features. Chen et al. [11] generated a few semantic tokens from the semantic representation of bitemporal images produced by CNNs and modeled the long-range contexts with a transformer encoder and decoder. After that, the ChangeFormer was introduced in [34], which dropped the CNNs and used a hierarchical transformer encoder with a multilayer perceptron decoder to render multiscale long-range dependencies. CD can be considered as a dense prediction task, where high-resolution contextual feature representations are always important [53]. However, the existing self-attention-based methods tend to focus extensively on the global context while neglecting the local contexts. The high memory requirement also limits their deployment on high-resolution feature maps, further reducing their effectiveness in capturing the accurate context.

To this end, we introduce an effective LGA that can efficiently perform short- and long-range visual dependencies between high-resolution inputs while suppressing irrelevant background information with a low memory requirement.

III. Proposed Method

This section presents the details of our method, including the overall network, memory-efficient LGA, mask-guided feature difference generation, and the loss function.

A. Network Overview

Fig. 2 gives an overview of the proposed MLA-Net, which is based on the Siamese network and consists of three components:

1) local–global attention network (LGA-Net); 2) mask-guided feature difference network (MFD-Net); and 3) prediction head. Formally, given the bitemporal RS images $\{X_1, X_2\} \in \mathbb{R}^{W\times H}$, the LGA-Net maps $\{X_1, X_2\}$ into a shared feature space to enhance their distinguishability. The LGA-Net comprises two branches with shared architecture and weights. Each branch includes a backbone network, an atrous spatial pyramid pooling (ASPP) [54] module, and a feature pyramid network (FPN) [55]. The backbone network extracts maps of different scales from bitemporal images, denoted as $\{F_1^i, F_2^i\}$. Here, $i \in \{1, 2, 3, 4\}$ indexes the feature extraction layers, with spatial resolution decreasing gradually from $\frac{W}{4} \times \frac{H}{4}$ to $\frac{W}{32} \times \frac{H}{32}$. Given the varying sizes of objects in bird's-eye long-distance imaging, we incorporate the ASPP module behind the backbone network to leverage its success in capturing multiscale contexts. The extracted features are then fed into the FPN to construct multiscale high-level semantic feature maps, denoted as $\{P_1^i, P_2^i\}_{\{i=1,2,3,4\}}$. To capture accurate contextual information, we introduce the proposed memory-efficient LGA in the first two layers of the backbone network.

The MFD-Net first performs feature subtraction between bitemporal images at different scales. The resulting feature maps then undergo a 3×3 convolution, batch normalization, and rectified linear unit layers to introduce nonlinearity, yielding the initial feature differences, i.e., $\{D^i\}_{\{i=1,2,3,4\}}$. Similar to the previous feature extraction step, D^4 is processed through the ASPP module to acquire multiscale change information. Subsequently, the change information at different scales is densely fused in a feedforward manner. Specifically, the ith scale is combined with the feature changes of all the previous scales $D^{i+1}, \ldots, D^4$ via

$$D^i = \text{fuse}\left(\text{Up}\left(\left[D^{i+1}, \ldots, D^4\right]\right), D^i\right) \tag{1}$$

where $\text{Up}(\cdot)$ denotes the upsampling operation and 1×1 convolution to align the feature maps in size, and $\text{fuse}(\cdot)$ is implemented with feature concatenation. To eliminate irrelevant changes, we introduced the change mask during the generation of feature differences.

The prediction head uses a 1×1 convolution $\text{conv}_{1\times 1}$, followed by a softmax function and an upsampling operation $\text{Up}(\cdot)$ to produce the change probability map, i.e.,

$$p = \text{softmax}\left(\text{Up}\left(\text{conv}_{1\times 1}(D)\right)\right) \tag{2}$$

where D is the concatenation of feature differences at different scales. We perform upsampling on different sets of features to ensure uniform scales and employ a 1×1 convolution to match the depths of the respective feature sets. In the testing process, p is binarized with a threshold of 0.5 to generate the final change map. In the next subsections, we introduce the memory-efficient LGA and mask-guided feature difference generation in detail.

B. Memory-Efficient LGA

RS objects have unclear details and appearance variance. Therefore, it is necessary to integrate the local and global interactions between different regions of images to improve the feature discriminative capability. Unfortunately, convolution usually learns local features related to edges and textures due to locally biased receptive fields. Self-attention-based global contextual information exploitation lacks the effective consideration of the local context and has a very high memory requirement, which hinders its feasibility for high-resolution RS images. To solve the problems, we develop a memory-efficient and effective LGA for local and global context exploitation.

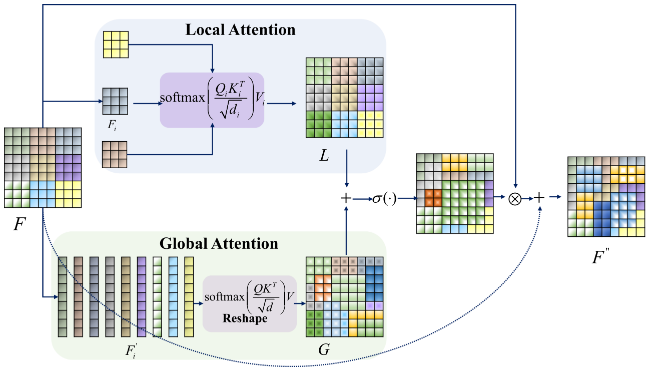

Local Attention
F_i
$\text{softmax}\left(\frac{Q_iK_i^T}{\sqrt{d_i}}\right)V_i$
L
F
$\sigma(\cdot)$
F''
Global Attention
F_i'
$\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$
Reshape
G

Fig. 3. Illustration of our LGA module.

Our method is based on the assumption that not all pixels have semantic dependencies and, therefore, the connections between pixels should be sparse rather than dense. The success of the nonlocal means algorithm [56], [57] suggests that there are long-range interactions between local patches that are spatially distant from each other. In other words, pixels in each local patch should be connected to characterize the local context, while all the patches should be connected to capture the global context. This observation drives us to introduce LGA to exploit this context.

As shown in Fig. 3, given a feature map $F \in \mathbb{R}^{W\times H\times C}$, we first divide it into multiple nonoverlapping patches of the same size $w \times h$, obtaining $F_i \in \mathbb{R}^{w\times h\times C}$. After that, self-attention within each patch is first computed to encode the local relationship between pixels, i.e.,

$$L_i\left(Q_i, K_i, V_i\right) = \text{softmax}\left(\frac{Q_iK_i^T}{\sqrt{d_i}}\right)V_i \tag{3}$$

where $\{Q_i, K_i, V_i\} \in \mathbb{R}^{wh\times C}$ are the query, key, and value, respectively, and are obtained by performing 1×1 convolutional operations on F_i and $d_i = C$. In this way, short-range spatial correlation is established, facilitating the extraction of image details. After that, the local attention of all patches is collected according to the spatial location to derive the local attention L of the whole feature map.

The global attention is constructed between patches. Specifically, feature maps within each patch are concatenated along the channel dimension. All the batches are rearranged to yield $F' = [F_1', \ldots, F_N']$ with $F_i' \in \mathbb{R}^{N\times Cwh}$ and N is the number of patches. Accordingly, self-attention between patches is calculated as follows:

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

where $\{Q, K, V\} \in \mathbb{R}^{N\times Cwh}$ are the query, key, and value, respectively, and are obtained by performing 1×1 convolutional operations on the input feature map F' and $d = Cwh$. This step allows the local patches to interact with each other, thus establishing global spatial semantic dependencies. Then, the obtained global attention is reshaped to the same size as the local attention. Thanks to the simultaneous use of local and global attention, the feature extraction ability of the network can be dramatically enhanced. Finally, the attentive features are obtained by

$$F'' = F + \sigma(\mathrm{L} + \mathrm{G}) \otimes F \tag{5}$$

where σ is the sigmoid function and $\otimes$ means the elementwise multiplication. This step further adaptively refines the feature maps at different locations based on local and global attentions, thus enabling the network to focus on the targets of interest.

Our LGA model can be considered as a combination of spatial attention and self-attention and has hybrid advantages but a low memory requirement. Removing the σ function and the elementwise multiplication in (5), the LGA becomes

$$F'' = F + L + G. \tag{6}$$

This is very similar to the scaled dot-product-based self-attention, except that our LGA module can effectively capture both local and global contextual information while suppressing the background information. The space complexity of our LGA module is $O(N(hw)^2)$, while that of self-attention is $O((WH)^2)$. As $Nhw = HW$, the space complexity of self-attention is N times that of the LGA module. Accordingly, our LGA can be flexibly deployed on high-resolution feature maps to enrich the contextual information extraction.

Moreover, when replacing the $L + G$ with performing convolutions on F, (5) reduces to

$$F'' = F + \sigma(\mathrm{conv}(F)) \otimes F. \tag{7}$$

This is the popular spatial attention. The difference is that our LGA can enjoy the benefits of the global contexts to enhance feature representation. In this way, our network is more capable of identifying changes of interest.

C. Mask-Guided Feature Difference Generation

Because of the imaging environment and background complexity, direct feature subtraction between bitemporal images tends to introduce irrelevant change information. We further introduce a change mask to refine the feature differences. This mask assigns higher importance to changed pixels and lower importance to unchanged ones. As shown in Fig. 4, the mask is generated by

$$\mathrm{Mask}^i = \delta\left(D^i\right) \tag{8}$$

where $\delta(\cdot)$ is the tanh function. As $\delta(\cdot)$ maps D^i into weight between 0 and 1, intuitively, (8) generates mask scores that can be used to choose positions related to real changes.

Embedding the mask into (1), we can obtain the masked feature fusion as follows:

$$D_f^i = \mathrm{fuse}\left(\mathrm{Mask}^i \otimes \left(\mathrm{Up}\left(\left[D^1, \ldots, D^{i-1}\right]\right), D^i\right)\right). \tag{9}$$

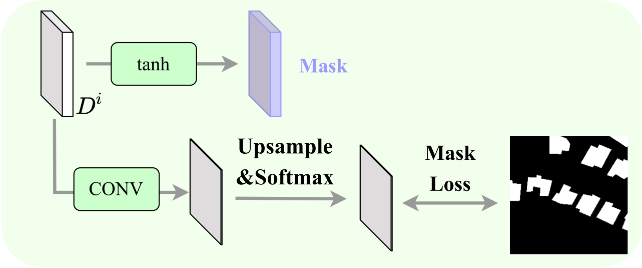


Fig. 4. Illustration of the change mask.

In this way, we obtain a hierarchical feature difference with higher accuracy, significantly overcoming the pseudo changes caused by the complex background.

In the absence of supervision, it is unclear whether the mask can force the network to extract the exact feature differences. Intuitively, a mask loss can be associated with the mask to guide the learning of the mask. However, the mask is the score that indicates the possibility of feature changes, and there are no corresponding ground truth labels to supervise the mask. Alternatively, as the mask is directly generated from D^i, the accurate D^i means a high-quality mask. Therefore, the mask loss can be imposed on D^i. The ground truth change label t is used to supervise D^i. For this regard, as shown in Fig. 4, we use a lightweight fully convolutional network with a sequence of convolution and upsampling operators and a softmax function to turn D^i into predicted change m. After that, the mask loss computes the binary cross-entropy (BCE) loss and the dice loss between m and the downsampled ground truth change map t, i.e.,

$$\mathcal{L}_{\text{Mask}} = L_{\text{Dice}}(m, t) + L_{\text{BCE}}(m, t). \tag{10}$$

Here, $L_{\text{Dice}}(m, t)$ is defined as

$$L_{\text{Dice}}(m, t) = \sum_{i=1}^{N} 1 - \frac{2m_i t_i}{m_i} \tag{11}$$

and $L_{\text{BCE}}(m, t)$ is

$$L_{\text{BCE}}(m, t) = -\sum_{i=1}^{N} t_i \log m_i + (1 - t_i) \log(1 - m_i) \tag{12}$$

where t_i and $1 - t_i$ denote the changed and unchanged pixels in the downsampled ground truth change map, respectively; N denotes the total number of pixels.

D. Loss Function

The MLA-Net contains three loss functions, i.e., BCE loss, dice loss, and mask loss:

$$L_{\text{total}} = L_{\text{BCE}} + L_{\text{Dice}} + \lambda \sum_i L_{\text{Mask}_i} \tag{13}$$

where λ balances the mask loss and the other two losses. We experimentally set $\lambda = 0.5$ as the best performance is achieved. The BCE loss calculates the loss between the ground truth change labels and the predicted labels. The dice loss accounts for the sample imbalance issue between the number of changed and unchanged regions.

IV. Experiments

We conducted experiments on the widely used LEVIR-CD, WHU-CD, and CLCD datasets to evaluate the performance of our method. Moreover, detailed analysis and ablation study are provided to show the effectiveness of our method.

A. Experiment Settings

1) Datasets: Three datasets are used for evaluation, including LEVIR-CD, WHU-CD, and CLCD.

a) The LEVIR-CD [46] dataset is on building CD and has 637 pairs of bitemporal images, each containing 1024 × 1024 pixels. According to the default settings in [46], each image was cut into small nonoverlapping patches in size of 256 × 256, producing 7120, 1024, and 2048 samples for training, validation, and testing, respectively.

b) The WHU-CD [58] dataset consists of one pair of aerial images of size 32 507 × 15 354. One contains 12 796 buildings captured in 2012, and the other contains 16 077 buildings captured in 2016. We cut this image into small nonoverlapping patches in size of 256 × 256. We randomly divided this dataset into 6096, 762, and 762 samples for training, validation, and testing, respectively.

c) The CLCD [59] dataset is a cropland change dataset and contains 600 pairs of images in the size of 512 × 512. It includes multiple changes related to the cropland, such as buildings, roads, lakes, and bare soil lands. Following the setting in [59], we split the dataset into 360, 120, and 120 samples for training, validation, and testing, respectively.

2) Evaluation Metrics: Five metrics were used to measure the CD performance, including precision (P), recall (R), F1 score (F1), intersection over union (IoU), and total precision (OA). These five metrics are defined as follows:

$$\begin{aligned}
\mathrm{P} &= \frac{\mathrm{TP}}{\mathrm{TP{+}FP}} \\
\mathrm{R} &= \frac{\mathrm{TP}}{\mathrm{TP{+}FN}} \\
\mathrm{IoU} &= \frac{\mathrm{TP}}{\mathrm{TP{+}FN{+}FP}} \\
\mathrm{OA} &= \frac{\mathrm{TP{+}TN}}{\mathrm{TP{+}TN{+}FN{+}FP}} \\
\mathrm{F1} &= \frac{2}{\mathrm{R}^{-1}+\mathrm{P}^{-1}}
\end{aligned} \tag{14}$$

where TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative, respectively.

3) Implementation Details: The patch size $w \times h$ was set as 8 × 8. The LGA modules are placed in the first and second layers of the FPN. The change masks are applied on $\{D^1, D^2\}$. We will examine their impact on the performance in the experiment. Our network was implemented using PyTorch and trained on NVIDIA GeForce RTX 3090 GPUs with the AdamW optimizer. The initial learning rate was set to 0.002. We used the OneCycleLR strategy to tune the learning rate with a maximum and minimum of 0.002 and 0.002/500, respectively. For both the LEVIR-CD and WHU-CD datasets, the batch size and the total number of epochs were set as 32 and 250, respectively. For the CLCD dataset, the batch size and the total number of epochs were 8 and 300, respectively.

TABLE I
Comparison of Different Methods on Results on the LEVIR-CD Dataset

Method		P	R	F1	IoU	OA
FC-EF		86.91	80.17	83.40	71.53	98.39
FC-Siam-Diff		89.53	83.31	86.31	75.92	98.67
FC-Siam-Conc		91.99	76.77	83.69	71.96	98.49
STANet		83.81	91.00	87.26	77.40	98.66
DTCDSCN		88.53	86.83	87.67	78.05	98.77
EGRCNN		88.58	**91.72**	90.12	82.03	98.97
ChangeFormer		92.05	88.80	90.40	82.48	99.04
BIT		89.24	89.37	89.31	80.68	98.92
ICIF-Net		91.39	89.24	90.38	82.31	98.99
MLA-Net	ResNet-18	91.69	90.06	90.87	83.27	99.08
	ResNet-50	$\underline{92.49}$	90.50	91.49	84.31	99.14
	ResNet-101	92.31	$\underline{91.04}$	$\underline{91.67}$	$\underline{84.62}$	$\underline{99.16}$
	EfficientNet-B4	**93.32**	90.16	**91.72**	**84.70**	**99.17**

Bold: best; Underline: second best.

4) Compared Methods: We selected nine methods for comparison, including CNN-based methods (FC-EF, FC-Siam-Diff, FC-Siam-Conc [36], STANet [46], DTCDSCN [60], and EGRCNN [16]), Transformer-based methods (BIT [11] and ChangeFormer [34]), and a hybrid CNN and Transformer-based method (ICIF-Net [7]). By default, we directly used the performance of the comparison methods reported in the original article. Otherwise, we trained them using their published code with the default parameters.

B. Comparisons With State of the Arts

1) Results on the LEVIR-CD Dataset: Table I gives quantitative comparisons of all the methods on the LEVIR-CD dataset. We employed different backbone networks to implement MLA-Net, including ResNet-18, ResNet-50, ResNet-101 [61], and EfficientNet-B4 [62], to test its robustness to feature extraction. Among all the alternative methods, ChangeFormer achieves the highest performance thanks to the inherent advantages in global spatial dependence modeling. Our MLA-Net provides very competitive performance even with ResNet-18 as the backbone network. The advanced backbone network can further improve the accuracy. These performance improvements are attributed to the strong ability of the introduced LGA module to capture the local–global context and the most important features successfully. This enhances the discriminative power of the network to identify the targeted object. In addition, the introduced change masks are more capable of filtering out irrelevant pseudo changes resulting from the complex background.

We further show the qualitative comparison on the LEVIR-CD dataset in Fig. 5. FC-EF, FC-Siam-Diff, and FC-Siam-Conc are removed because of their relatively poor performance. To provide a better illustration, the true positive and true negative are marked white and black, respectively. The false positive and false negative are marked green and red, respectively. The visual results of our MLA-Net are presented by setting the ResNet-18 as the backbone. Although the buildings are overwhelmed by strong lighting and shadows with appearance diversity, our

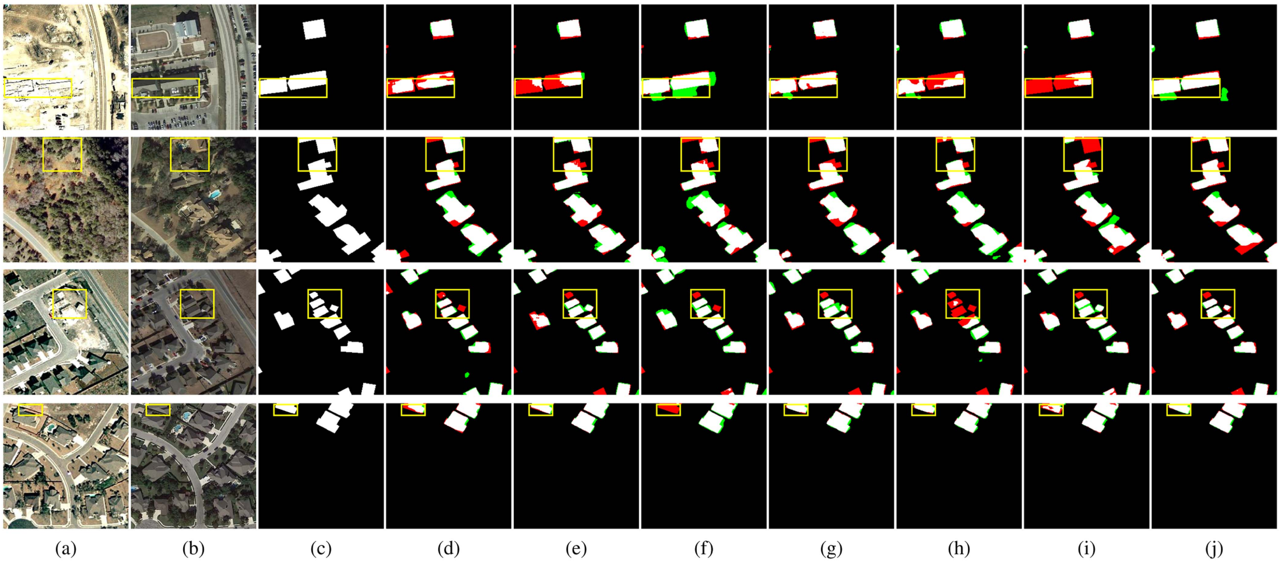


Fig. 5. Visual comparison on the LEVIR-CD dataset. (a) T1 images. (b) T2 images. (c) Ground truth. (d) BIT. (e) ChangeFormer. (f) EGRCNN. (g) DTCDSCN. (h) STANet. (i) ICIF-Net. (j) Ours.

TABLE II
COMPARISON OF DIFFERENT METHODS ON THE MEMORY, NUMBER OF PARAMETERS, TRAINING TIME, AND TESTING TIME

Method	Memory	#Params	Training	Testing
FC-EF	5147 MiB	1.351 M	38.81 s	7.49 s
FC-Siam-Diff	5601 MiB	1.350 M	46.13 s	9.72 s
FC-Siam-Conc	5781 MiB	1.546 M	49.17 s	10.05 s
STANet	31 980 MiB	16.892 M	226.62 s	33.31 s
DTCDSCN	5121 MiB	31.257 M	77.37 s	11.81 s
EGRCNN	8015 MiB	9.632 M	195.52 s	28.38 s
ChangeFormer	17 813 MiB	41.027 M	347.21 s	53.64 s
BIT	10 029 MiB	3.496 M	99.47 s	15.57 s
ICIF-Net	13 843 MiB	23.843 M	279.50 s	40.10 s
Ours	6180 MiB	176.942 M	134.49 s	30.56 s

TABLE III
COMPARISON OF DIFFERENT METHODS ON THE WHU-CD DATASET

Method		P	R	F1	IoU	OA
FC-EF		83.54	73.85	78.39	64.47	98.21
FC-Siam-Diff		85.92	78.89	82.26	69.86	98.50
FC-Siam-Conc		82.46	85.24	83.83	72.16	98.55
STANet		85.10	79.40	82.20	69.70	98.50
DTCDSCN		91.42	87.60	89.47	80.94	99.09
EGRCNN		92.60	86.55	89.47	80.04	99.10
ChangeFormer		92.06	83.46	87.55	77.86	98.96
BIT		93.91	87.84	90.78	83.11	99.21
ICIF-Net		91.19	85.92	88.48	79.34	99.01
MLA-Net	ResNet-18	96.14	92.11	94.08	88.83	99.49
	ResNet-50	95.85	92.72	94.26	89.14	99.50
	ResNet-101	94.77	**95.10**	94.93	90.35	99.55
	EfficientNet-B4	**96.72**	93.45	**95.06**	**90.58**	**99.57**

Bold: best; Underline: second best.

LGA-Net can more effectively extract accurate change masks with higher completeness and fewer false detections and misdetections. It can also be observed that some buildings are difficult to distinguish due to the appearance diversity and the occlusion of trees. As a result, BIT, ChangeFormer, and STANet generate more false negatives. In contrast, the LGA module helps our MLA-Net successfully detect the building changes by establishing local and global contexts and refining the feature maps spatially. On the other hand, shadows, noise, and strong light can also introduce false changes. Due to the advanced feature difference extraction enabled by the introduced change masks, our MLA-Net is more capable of identifying the true changes and presents fewer false positives and better completeness.

In Table II, we present a comprehensive comparison of other aspects, including memory consumption, parameters, as well as training and testing time, under consistent settings. For example, the batch size is 16 and input image size is 256 × 256. Here, we use ResNet-18 as the backbone network to implement our MLA-Net. Notably, our method stands out by occupying relatively less memory while achieving the highest performance among all considered methods. This observation underscores the positive impact of the LGA and change mask in our approach. It is worth noting that, in comparison to alternative methods, our approach exhibits fewer advantages in terms of parameter count, training time, and testing time. However, striking a balance among these factors is inherently challenging in the development of a method. Acknowledging this difficulty, we view the optimization of these aspects as our future work.

2) Results on the WHU-CD Dataset: Here, we report the results on the WHU-CD dataset. To eliminate the impact of data splitting, we retrained all the competing methods. As presented in Table III, our method stands out and achieves higher scores in F1, IoU, and OA metrics. Compared to the second best method, BIT, our MLA-Net has a gain of 3.3% in the F1 metric when using the same backbone ResNet-18. The performance improvement is mainly attributed to the introduced change mask, which effectively suppresses false changes caused by the surrounding background.

Fig. 6 shows the qualitative comparison on the WHU-CD dataset. Buildings and other objects, such as containers and vehicles, have very similar visual appearances, leading to low

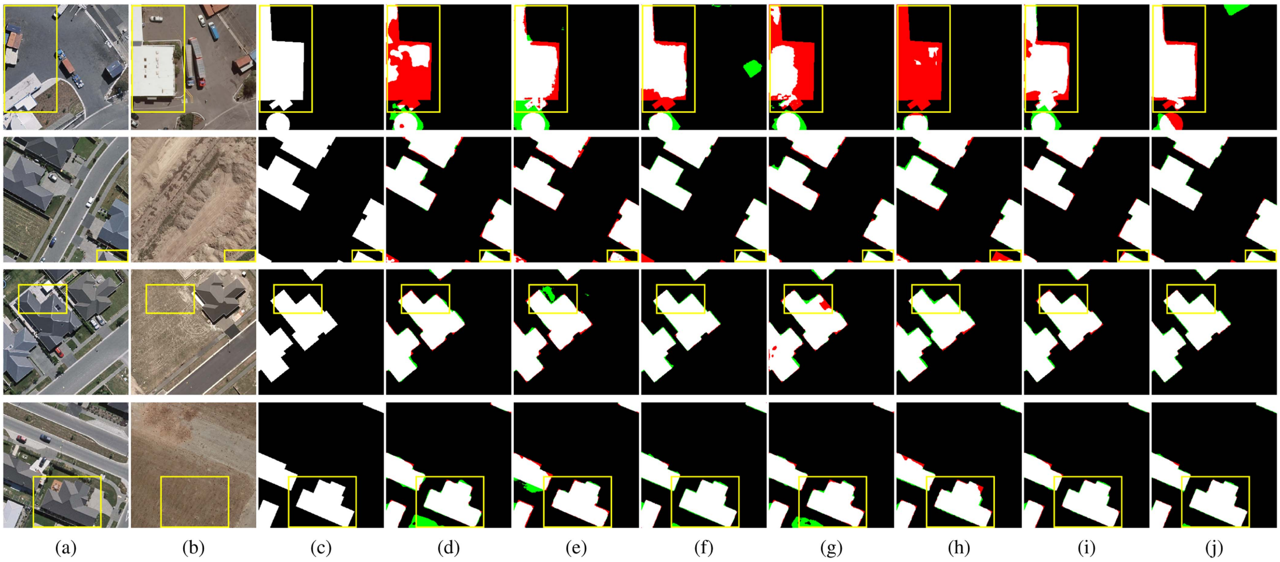


Fig. 6. Visual comparison on the WHU-CD dataset. (a) T1 images. (b) T2 images. (c) Ground truth. (d) BIT. (e) ChangeFormer. (f) EGRCNN. (g) DTCDSCN. (h) STANet. (i) ICIF-Net. (j) Ours.

variation between classes. The LGA module improves the discriminative power of the network thanks to the simultaneous exploitation of local and global contexts and the refinement of spatial feature maps. As a result, our MLA-Net can more adequately identify building changes and deliver noticeably fewer false negatives. The edge guidance module in EGRCNN helps to detect buildings. For this reason, it obtains visually preferable performance by presenting more accurate change boundaries. It can also be seen that our MLA-Net is more robust to pseudo changes compared to EGRCNN. The main reason is that more reliable feature differences can be extracted due to the introduction of the change masks. The consistently excellent performance on the WHU-CD dataset strongly suggests the effectiveness of our method for CD.

3) Results on the CLCD Dataset: We further evaluated all the methods on the CLCD dataset. Unlike the LEVIR-CD and WHU-CD datasets, which focus only on building changes, the CLCD dataset focuses on changes related to cultivated land. This may be caused by multiple objects, such as roads and lakes, making changes more difficult to distinguish. As a result, all the methods show a performance drop on this dataset, as shown in Table IV. The irregularity and unclear boundaries of roads and rivers make it troublesome to learn the edge detection module. Due to this reason, the performance of EGRCNN is poor. It should be noted that our method significantly outperforms alternative methods. The noticeable performance improvements are attributed to the following two factors. Armed with memory-efficient LGA module, our method utilizes both local and global contexts while effectively filtering out irrelevant information, facilitating extracting accurate changes. In addition, the mask-guided feature difference generation proves advantageous in suppressing pseudo and unwanted changes, leading to more accurate and refined results.

We provide visual results of all the methods in Fig. 7. In the first scene, the spatial resolution is very low, and the imaging condition also changes between bitemporal images, threatening challenges for distinguishing boundaries. Suffering from limited consideration of local information, most methods fail to detect the changes caused by the buildings. Attributing to the rich local–global contextual information powered by the introduced LGA module, our method is able to extract most of the changes. In the next two scenes, the roads are more challenging to be told from the surrounding environment, which inevitably introduces more false positives and false negatives when detecting changes. However, our model still adapts well to changes in road contours due to the powerful contextual modeling capability of the introduced LGA module. The last two scenes contain multiple changes, including hills and buildings. Not all are associated with cropland, making it very demanding for the model to extract the changes of interest accurately. Benefiting from the hybrid merits of the LGA module and the change mask, our method can effectively reduce the impact of irrelevant changes and accurately detect cropland changes. This is not the case with other methods. Overall, this experiment demonstrates the effectiveness of our method in very complicated scenarios.

TABLE IV
COMPARISON OF DIFFERENT METHODS ON THE CLCD DATASET

Method		P	R	F1	IoU	OA
FC-EF		58.88	56.32	57.57	40.42	93.82
FC-Siam-Diff		59.27	62.38	60.79	43,66	94.31
FC-Siam-Conc		61.71	65.29	63.00	45.99	94.85
STANet		55.80	68.00	61.30	38.40	93.60
DTCDSCN		61.25	59.11	60.16	43.02	94.18
EGRCNN		67.12	69.04	68.07	51.59	95.18
ChangeFormer		58.29	47.25	52.19	35.31	94.48
BIT		64.18	58.63	61.28	44.18	94.48
ICIF-Net		66.84	54.02	58.75	42.60	94.58
MLA-Net	ResNet-18	78.56	75.66	77.08	62.71	96.65
	ResNet-50	80.58	73.42	76.83	62.38	96.71
	ResNet-101	**83.17**	76.93	79.87	66.49	**97.12**
	EfficientNet-B4	80.80	**78.77**	**79.88**	**66.50**	97.04

Bold: best; Underline: second best.

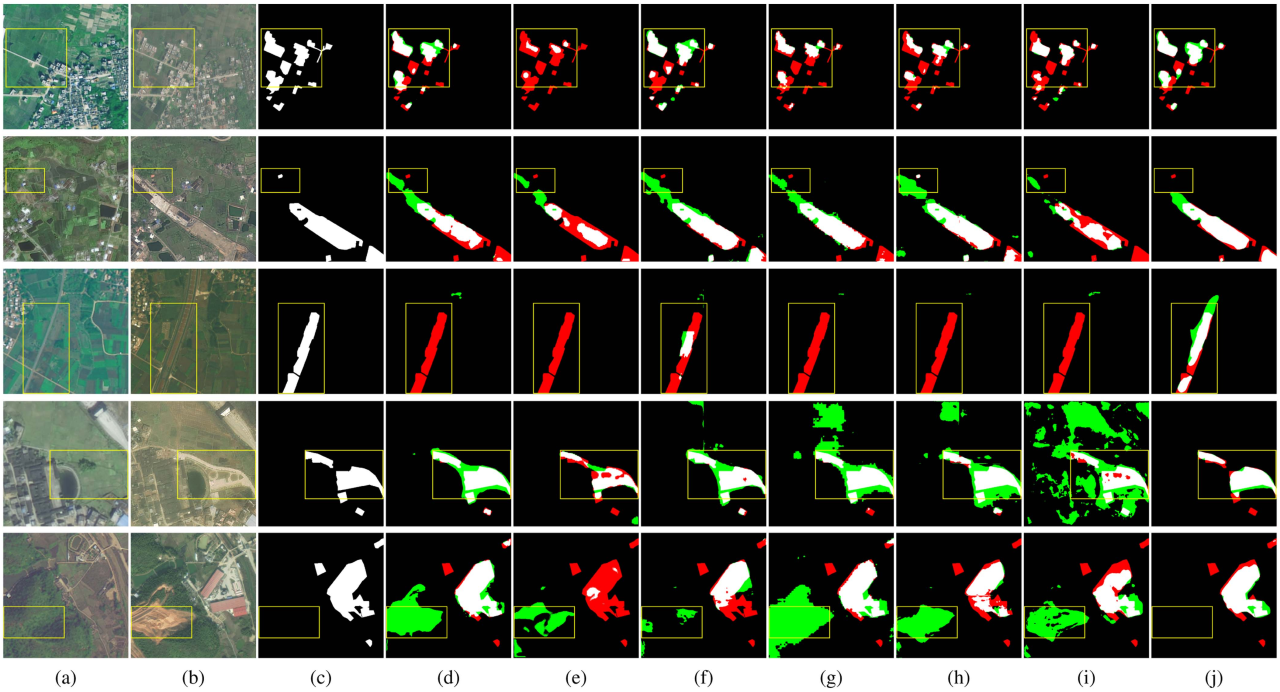


Fig. 7. Visual comparison on the CLCD dataset. (a) T1 images. (b) T2 images. (c) Ground truth. (d) BIT. (e) ChangeFormer. (f) EGRCNN. (g) DTCDSCN. (h) STANet. (i) ICIF-Net. (j) Ours.

TABLE V
Ablation Study on Different Modules

Baseline	LGA	Change Mask	Mask Loss	LEVIR-CD P/R/F1/IoU/OA	WHU-CD P/R/F1/IoU/OA	CLCD P/R/F1/IoU/OA
✓	✗	✗	✗	88.21/89.51/88.86/79.96/98.86	95.31/87.62/91.30/84.00/99.27	79.00/70.64/74.59/59.47/96.42
✓	✓	✗	✗	89.54/89.89/89.72/81.35/98.95	96.44/88.89/92.51/86.07/99.37	77.95/73.95/75.90/61.16/96.50
✓	✓	✓	✗	89.02/**90.92**/89.96/81.75/98.97	**96.52**/89.39/92.82/86.60/99.39	**82.14**/71.24/76.31/61.69/**96.71**
✓	✓	✗	✓	90.05/90.14/90.09/81.97/98.99	94.06/**92.17**/93.10/87.10/99.40	76.12/**75.86**/75.99/61.28/96.43
✓	✓	✓	✓	**91.69**/90.06/**90.87/83.27/99.08**	96.14/92.11/**94.08/88.83/99.49**	78.56/75.66/**77.08/62.71**/96.65

The bold values indicate the best performance.

C. Ablation Study

To gain a more comprehensive understanding of our MLA-Net, we performed an ablation study of the different components in this section. Except the first one, all the experiments were performed on the LEVIR-CD dataset.

1) Effectiveness of Introduced Modules: We built the baseline model by removing the change mask, mask loss, and LGA modules and selecting ResNet-18 as the backbone network. LGA identifies the most informative locations on a local–global scale. Thus, as given in Table V, performance improves with the LGA module. The additional performance gain can be obtained using masks and corresponding losses to filter out irrelevant pseudo feature differences. It is worth noting that the combination of three beneficial components achieves the most remarkable performance gain, as indicated by the F1 score.

To visually demonstrate the effectiveness of the change mask and loss, we present the feature differences extracted from the last four cases in Fig. 8. It can be seen that more accurate feature differences can be obtained by suppressing irrelevant ones and highlighting the corrected ones with the help of the change mask and corresponding loss. For clearer intuition, we also present the obtained change mask in Fig. 9. It is evident that the additional supervision results in a precise mask, which more closely resembles the ground truth change. Overall, the ablation study evidently verifies the effectiveness of the introduced LGA module and change mask in improving CD accuracy.

TABLE VI
Comparison With Alternative Attentions

Attentions	P	R	F1	IoU	OA	FLOPS	Memory
Self-attention	**91.83**	89.07	90.43	82.53	99.04	14.111G	28.408G
Spatial-attention	90.71	89.19	89.94	81.72	98.98	**13.654G**	12.537G
LGA **(Ours)**	91.69	**90.06**	**90.87**	**83.27**	**99.08**	29.630G	**8.925G**

The bold values indicate the best performance.

2) Comparison With Alternative Attentions: To fully demonstrate the effectiveness of the proposed LGA module, we further replaced the LGA module with self-attention in (6) and spatial attention in (7). Equation (7) was implemented with spatial attention in the convolutional block attention module [63]. As can be seen from Table VI, the self-attention mechanism tends to focus extensively on the global information but overlooks the local contexts, leading to lower CD accuracy. The high memory consumption also limits its usage in high-resolution feature maps. Our LGA provides the best performance, which

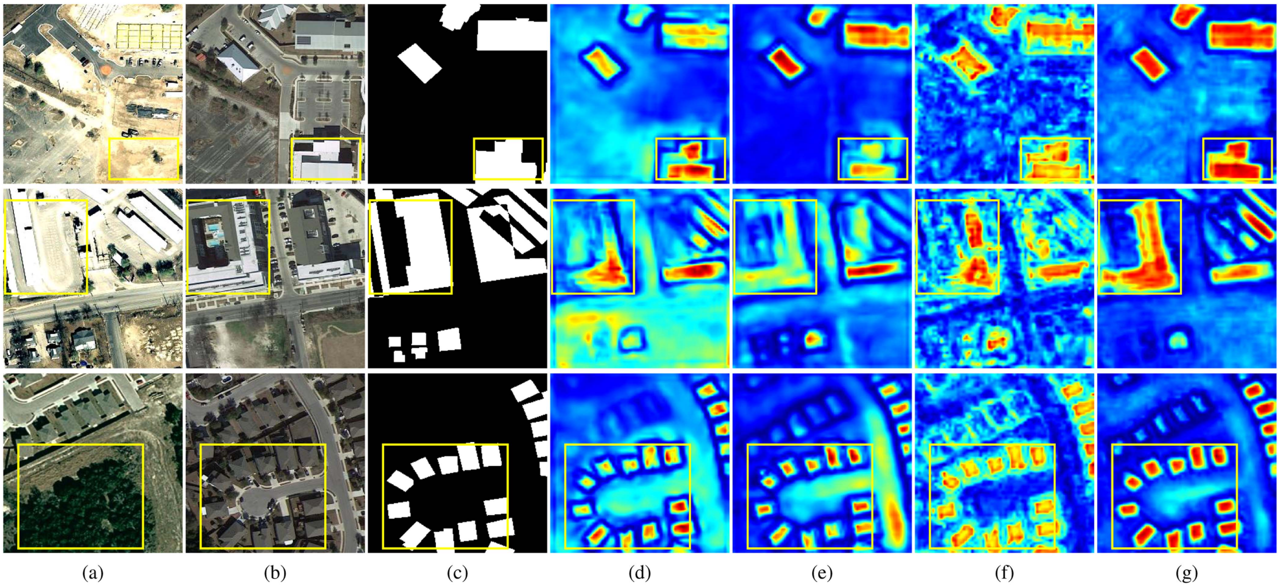


Fig. 8. Visualization of extracted feature differences. (a)–(c) Images and ground truth changes. (d)–(g) Feature differences yielded by Baseline+LGA, Baseline+LGA+Mask, Baseline+LGA+Loss, and Baseline+LGA+Mask+Loss.

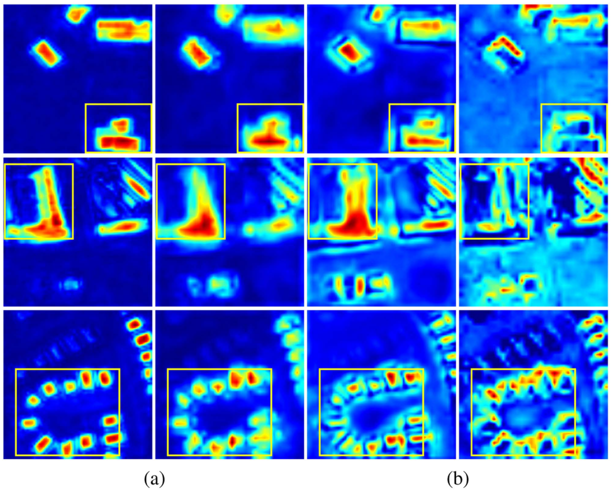


Fig. 9. Impact of the additional supervision on the change mask generation. (a) With. (b) Without.

TABLE VII
ABLATION STUDY ON L AND G IN THE LGA MODULE

L	G	P	R	F1	Iou	OA
✓	✗	**92.14**	89.12	90.61	82.83	99.05
✗	✓	90.34	**90.65**	90.49	82.64	99.03
✓	✓	91.69	90.06	**90.87**	**83.27**	**99.08**

The bold values indicate the best performance.

TABLE VIII
IMPACT OF THE PATCH SIZE ON THE LGA MODULE

Size	P	R	F1	IoU	OA
1 × 1	**91.70**	89.38	90.53	82.70	99.04
2 × 2	89.78	89.19	89.48	80.97	98.93
4 × 4	90.23	**90.59**	90.41	82.50	99.02
8 × 8	91.69	90.06	**90.87**	**83.27**	**99.08**
16 × 16	91.60	89.96	90.77	83.10	99.06

The bold values indicate the best performance.

is attributed to the hybrid advantage of spatial attention and self-attention. We acknowledge that our method involves a larger computational complexity than some previous works such as self-attention and spatial attention. However, we believe that it is worth paying the additional computational cost for improved performance. Moreover, our LGA module consumes less memory, making it more feasible for implementation on satellites with limited memory space.

To gain further intuition, we visualize the resulted feature maps in Fig. 10. It can be clearly seen that the attention generated by our LGA module makes the network focus on the most important locations related to buildings. This also justifies the superior performance in Table VI.

We further conducted an ablation study on the local and global attention in the LGA module to show their contribution to CD. As indicated in Table VII, both L and G exhibit discernible impacts on the performance, and their combined utilization contributes to an overall higher performance.

3) Patch Size of the LGA Module: The size of the patch controls the memory consumption and the effectiveness of LGA. Small sizes require a high memory load and result in local information not being used effectively. Large size may cause the global context to be underutilized. Table VIII presents the performance of LGA peaks at 8 × 8, balancing local and global context exploitation.

4) Locations of LGA Modules: By setting the patch size to 8 × 8, we change the LGA module in different layers of the FPN, from layer 1 to layer 3, to observe the performance change. Accordingly, the number of LGA modules increases from 1 to 3. As shown in Table IX, the best performance is obtained by setting the module in the first two layers, corresponding to feature maps

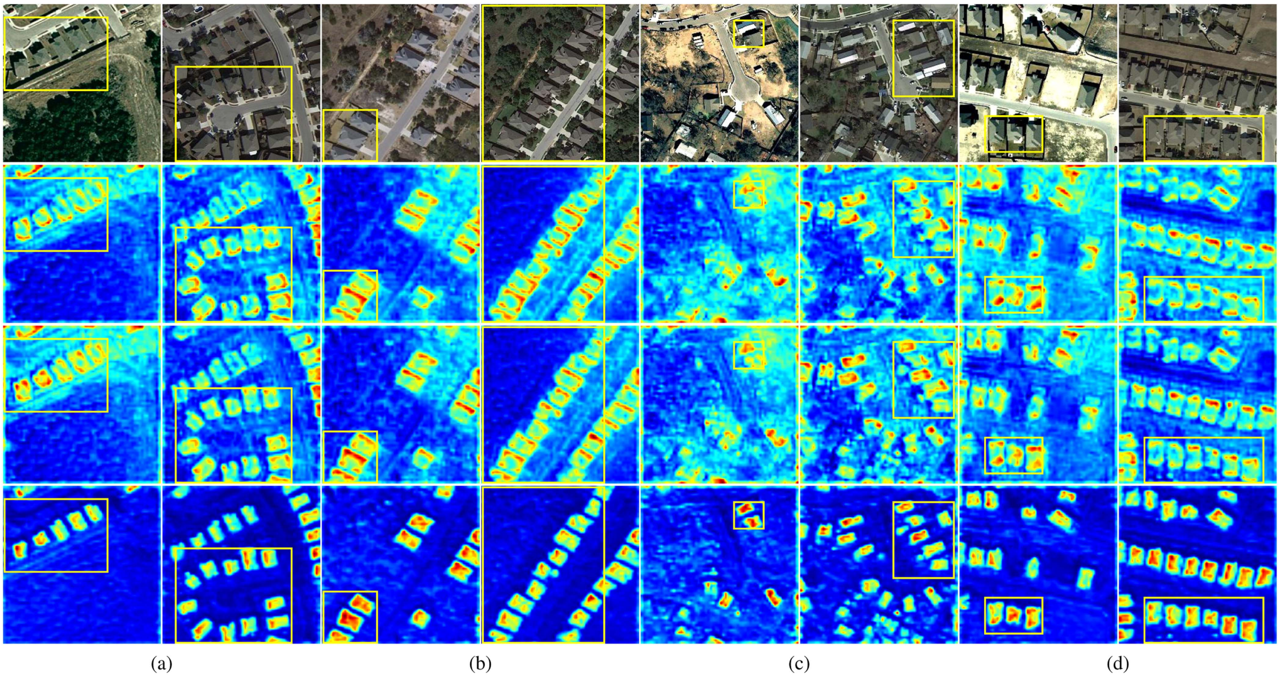


Fig. 10. Visualization of the feature maps generated by the different attention modules. From top to bottom are the captured images, the feature maps of the spatial attention, self-attention, and LGA modules.

TABLE IX
IMPACT OF THE LOCATIONS OF LGA MODULES

First	Second	Third	P	R	F1	Iou	OA
✓	✗	✗	**92.94**	88.70	90.77	83.10	**99.08**
✓	✓	✗	91.69	**90.06**	**90.87**	**83.27**	**99.08**
✓	✓	✓	92.79	88.62	90.65	82.91	99.07

The bold values indicate the best performance.

TABLE X
IMPACT OF THE LOCATIONS OF CHANGE MASKS

D^1	D^2	D^3	D^4	P	R	F1	IoU	OA
✓	✗	✗	✗	**92.40**	88.59	90.46	82.58	99.04
✓	✓	✗	✗	91.69	**90.06**	**90.87**	**83.27**	**99.08**
✓	✓	✓	✗	91.85	88.64	90.22	82.18	99.02
✓	✓	✓	✓	91.34	89.04	90.18	82.11	99.01

The bold values indicate the best performance.

with higher resolution. In this case, the local characteristics can be more adequately captured while requiring fewer parameters.

5) Locations of Change Masks: As the experiment with the setting of LGA modules, we further vary the locations of masks to see how they affect the performance. From Table X, the highest performance is obtained when adding masks to $\{D^1, D^2\}$. We gradually fuse the feature differences from the lowest to the highest resolution. Feature maps in $\{D^1, D^2\}$ actually contain finer spatial information about the changes. Therefore, more accurate change masks can be generated, dramatically alleviating pseudo changes.

V. DISCUSSION

RS CD confronts challenges arising from the high diversity in target appearances and the complexity of backgrounds. Our article addresses the first challenge by introducing a memory-efficient LGA mechanism, which effectively extracts discriminative contextual information. The second challenge is mitigated through the incorporation of a change mask, guiding the generation of feature differences to identify genuine changes of interest. Moreover, the feature refinement in the LGA module is also helpful to tackle this issue. Results from both quantitative and qualitative experiments demonstrate that our approach achieves superior performance while maintaining low memory consumption.

However, our method has certain limitations. First, it requires a substantial number of parameters. Addressing the challenge of reducing the parameter number while preserving performance will be a key focus of our future research. Second, our approach does not incorporate prior knowledge of object shapes; instead, it relies solely on data-driven methods to achieve a highly discriminative representation. Improving the robustness of the method could be achieved by embedding object shape information into the network, perhaps in the form of a dedicated module or loss function.

VI. CONCLUSION

This article presents a novel MLA-Net to improve CD performance in RS images. The memory-efficient LGA module leverages the hybrid advantages of self-attention and spatial attention to acquire more discriminative context information. The change mask can extract more accurate feature differences, greatly suppressing the interference of irrelevant changes. Our MLA-Net achieves superior performance on three benchmark datasets, demonstrating the effectiveness of our approach in RS image CD.

References

[1] G. Reiersen et al., "ReforesTree: A dataset for estimating tropical forest carbon stock with deep learning and aerial imagery," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 12119–12125.

[2] J. Wang et al., "SSCFNet: A spatial-spectral cross fusion network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4000–4012, 2023.

[3] K. Yang et al., "Asymmetric Siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609818.

[4] L. Bergamasco, S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised change detection using convolutional-autoencoder multiresolution features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408119.

[5] J. Shi, T. Wu, A. K. Qin, Y. Lei, and G. Jeon, "Semisupervised adaptive ladder network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408220.

[6] 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.

[7] 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, 2022, Art. no. 4410213.

[8] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, 2023.

[9] D. Hong et al., "SpectraLGPT: Spectral foundation model," 2023, *arXiv:2311.07113*.

[10] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[11] 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.

[12] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, "Deep multiscale Siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406512.

[13] Y. Shangguan, J. Li, and L. Chang, "Dual-attention cross fusion context network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8943–8959, 2023.

[14] X. Zhang, W. Yu, and M.-O. Pun, "Multilevel deformable attention-aggregated networks for change detection in bitemporal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621518.

[15] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623811.

[16] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610613.

[17] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

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

[19] L. Zhang, X. Hu, M. Zhang, Z. Shu, and H. Zhou, "Object-level change detection with a dual correlation attention-guided detector," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 147–160, 2021.

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

[21] J. Zheng et al., "MDESNet: Multitask difference-enhanced Siamese network for building change detection in high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 15, Art. no. 3775, 2022.

[22] J. Deng, K. Wang, Y. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, 2008.

[23] M. Zanetti, F. Bovolo, and L. Bruzzone, "Rayleigh-rice mixture parameter estimation via EM algorithm for change detection in multispectral images," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5004–5016, Dec. 2015.

[24] L. Bruzzone and D. Prieto, "An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 452–466, Apr. 2002.

[25] B. Du, Y. Wang, C. Wu, and L. Zhang, "Unsupervised scene change detection via latent Dirichlet allocation and multivariate alteration detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4676–4689, Dec. 2018.

[26] Y. Li, M. Gong, L. Jiao, L. Li, and R. Stolkin, "Change-detection map learning using matching pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4712–4723, Aug. 2015.

[27] H. Chen, W. Li, S. Chen, and Z. Shi, "Semantic-aware dense representation learning for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630018.

[28] 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.

[29] J. Wang, Y. Zhong, and L. Zhang, "Change detection based on supervised contrastive learning for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601816.

[30] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, Aug. 2023.

[31] G. Liu, L. Li, L. Jiao, Y. Dong, and X. Li, "Stacked fisher autoencoder for SAR change detection," *Pattern Recognit.*, vol. 96, 2019, Art. no. 106971.

[32] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[33] M. Gong, X. Niu, P. Zhang, and Z. Li, "Generative adversarial networks for change detection in multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2310–2314, Dec. 2017.

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

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

[36] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[38] J. Li, S. Li, and F. Wang, "Adaptive fusion nestedUNet for change detection using optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5374–5386, 2023.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[40] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[41] 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.

[42] T. Liu et al., "Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4704817.

[43] 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.

[44] 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, 2021, Art. no. 5406216.

[45] W. Gao, Y. Sun, X. Han, Y. Zhang, L. Zhang, and Y. Hu, "AMIO-Net: An attention-based multiscale input–output network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2079–2093, 2023.

[46] 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.

[47] 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.

[48] 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, 2022, Art. no. 5604816.

[49] 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.

[50] X. Xu, J. Li, and Z. Chen, "TCIANet: Transformer-based context information aggregation network for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1951–1971, 2023.

[51] Q. Guo, X. Qiu, P. Liu, X. Xue, and Z. Zhang, "Multi-scale self-attention for text classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7847–7854.

[52] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, 2022.

[53] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[54] F. S. H. A. Liang-Chieh Chen and G. Papandreou, "Rethinking atrous convolution for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[55] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 239–256.

[56] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Process. On Line*, vol. 1, pp. 208–212, 2011.

[57] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 60–65.

[58] 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.

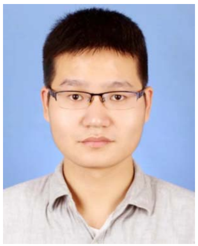
[59] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[60] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Trans. Geosci. Remote Sens.*, vol. 18, no. 5, pp. 811–815, May 2021.

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

[62] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[63] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.



Fengchao Xiong (Member, IEEE) received the B.E. degree in software engineering from Shandong University, Jinan, China, in 2014, and the Ph.D. degree in computer science and technology from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2019.

He visited Wuhan University, Wuhan, China; Griffith University, Nathan, QLD, Australia; and the University of Macau, Taipa, Macau, in 2011–2012, 2017–2018, and 2021–2023, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include hyperspectral image processing, machine learning, and pattern recognition.

Dr. Xiong is a Topical Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Tianhan Li received the B.E. degree in computer science and technology in 2021 from the Nanjing University of Science and Technology, Nanjing, China, where he is currently working toward the M.S. degree.

His research interests include machine learning and remote sensing image analysis.

Jingzhou Chen received the B.E. degree in computer science and technology from Sichuan University, Sichuan, China, in 2016, and the Ph.D. degree in computer science and technology from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2022.

From 2022 to 2023, he was a Senior Computer Vision Engineer with the Security and Risk Management Group, Ant Group, Hangzhou, China, where his work aimed to manage the risk from the image content. He is currently an Assistant Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include machine learning, pattern recognition, and remote sensing image analysis.

Jun Zhou (Senior Member, IEEE) received the B.S. degree in computer science and the B.E. degree in international business from the Nanjing University of Science and Technology, Nanjing, China, in 1996 and 1998, respectively, the M.S. degree in computer science from Concordia University, Montreal, QC, Canada, in 2002, and the Ph.D. degree in computing science from the University of Alberta, Edmonton, AB, Canada, in 2006.

In 2012, he joined the School of Information and Communication Technology, Griffith University, Nathan, QLD, Australia, where he is currently a Professor. Prior to this, he was a Research Fellow with the Research School of Computer Science, Australian National University, Canberra, ACT, Australia, and a Researcher with the Canberra Research Laboratory, NICTA, Canberra. His research interests include pattern recognition, computer vision, and spectral imaging with their applications in remote sensing and environmental informatics.

Dr. Zhou is an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *Pattern Recognition*.

Yuntao Qian (Senior Member, IEEE) received the B.E. and M.E. degrees in automatic control from Xi'an Jiaotong University, Xi'an, China, in 1989 and 1992, respectively, and the Ph.D. degree in signal processing from Xidian University, Xi'an, in 1996.

From 1996 to 1998, he was a Postdoctoral Fellow with Northwestern Polytechnical University, Xi'an. Since 1998, he has been with the College of Computer Science, Zhejiang University, Hangzhou, China, where he became a Professor in 2002. In 1999–2001, 2006, 2010, 2013, 2015–2016, and 2018, he was a Visiting Professor with Concordia University, Montreal, QC, Canada; Hong Kong Baptist University, Hong Kong; Carnegie Mellon University, Pittsburgh, PA, USA; the Canberra Research Laboratory, NICTA, Canberra, ACT, Australia; Macau University, Taipa, Macau; and Griffith University, Nathan, QLD, Australia. His current research interests include machine learning, signal and image processing, pattern recognition, and hyperspectral imaging.

Dr. Qian is an Associate Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.