SRC-Net: Bitemporal Spatial Relationship Concerned Network for Change Detection

Hongjia Chen , Xin Xu , and Fangling Pu , *Member, IEEE*

Abstract—Change detection (CD) in remote sensing imagery is a crucial task with applications in environmental monitoring, urban development, and disaster management. CD involves utilizing bitemporal images to identify changes over time. The bitemporal spatial relationships between features at the same location at different times play a key role in this process. However, existing change detection networks often do not fully leverage these spatial relationships during bitemporal feature extraction and fusion. In this work, we propose SRC-Net: a bitemporal spatial relationship concerned network for CD. The proposed SRC-Net includes a perception and interaction module that incorporates spatial relationships and establishes a cross-branch perception mechanism to enhance the precision and robustness of feature extraction. In addition, a patch–mode joint feature fusion module is introduced to address information loss in current methods. It considers different change modes and concerns about spatial relationships, resulting in more expressive fusion features. Furthermore, we construct a novel network using these two relationship concerned modules and conducted experiments on the LEVIR-CD and WHU Building datasets. The experimental results demonstrate that our network outperforms state-of-the-art methods while maintaining a modest parameter count. We believe that our approach sets a new paradigm for CD and will inspire further advancements in the field.

Index Terms—Change detection (CD), deep learning, optical remote sensing images, spatial relationship.

I. INTRODUCTION

CHANGE detection (CD) aims to identify change areas in the study region through analysis of multiple remote sensing observations [1], and has gradually become a research hotspot in remote sensing [2], [3]. By comparing remote sensing images taken at different times of the same region, change areas can be extracted on a pixel-by-pixel map. CD has played a significant role in many fields, including environment monitoring [4], [5], [6], resource management [7], urban planning [8], and disaster assessment [9], [10].

In recent years, with the widespread adoption of deep learning, particularly the advancements in convolutional neural networks (CNNs) [11] and Transformers [12], CD has gradually transitioned from traditional pixel-based and object-based methods [13], [14], [15], [16] to two-stage deep learning approaches, and finally to the end-to-end deep learning networks [17], [18]. However, most existing CD networks have originated from image semantic segmentation, lacking structure design for perceiving bitemporal spatial relationships. Therefore, existing methods are insufficient in the utilization of bitemporal spatial relationships, which is crucial for CD.

In this article, we define the corresponding relationships that exist between features at the same location at different times as bitemporal spatial relationships. The optical characteristics obtained from imaging the same location at different times should exhibit corresponding relationships. Similarly, the depth features extracted from the aforementioned optical characteristics should also exhibit corresponding relationships. Current network structures are rarely designed to take advantage of bitemporal spatial relationships. This is similar to using fully connected neural networks (FCs) to analyze images before CNNs were proposed. Fortunately, networks could perceive partial bitemporal spatial relationships with extensive learning. This is similar to how FCs can also perceive some spatial information in images, despite the difficulty involved.

Existing network structures can be categorized into two groups: single-branch networks and dual-branch networks. In single-branch networks, the bitemporal inputs are concatenated into one matrix as distinct channels, which enables semantic segmentation networks to address the CD task. In dual-branch networks, the bitemporal inputs are processed through a Siamese network with shared network parameters to extract distinct feature maps for each temporal, subsequently fusing these feature maps for the change area prediction. Intuitively, CD should analyze the ground object features at each time, and then compare the corresponding features between different times to detect changes. This is similar to the process of dual-branch networks that first extract features from the bitemporal inputs separately and then compare the differences between the bitemporal feature maps. Besides, the connection operations within a single-branch network lead to the disappearance of correspondence relationships among bitemporal features, thereby losing some bitemporal spatial relationships. This inevitably gives the impression that dual-branch networks should have advantage over single-branch networks. However, in the early-used dual-branch networks [19], each branch extracts features independently, and there is no cross-branch interaction. These networks did not perform well in terms of actual results and were significantly
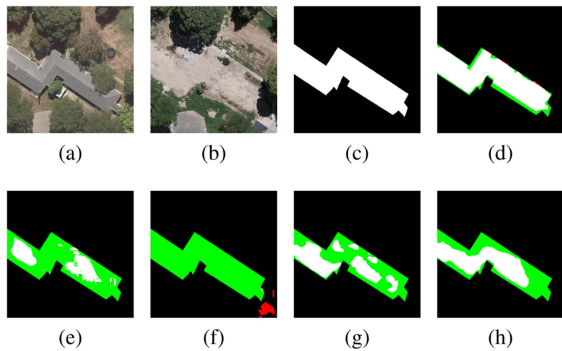
Fig. 1.   Bitemporal remote sensing images. (a) and (b) Original images. (c) Ground truth. (d) Result of our SRC-Net. (e) Result of the single-branch network (RDPNet). (f) Result of the dual-branch network without cross-branch interaction (FC-Siam-diff). (g) and (h) Results of dual-branch network with cross-branch interaction (BIT, DSIFN). The false positives and false negatives are indicated in red and green, respectively. Other colors represent true positives.
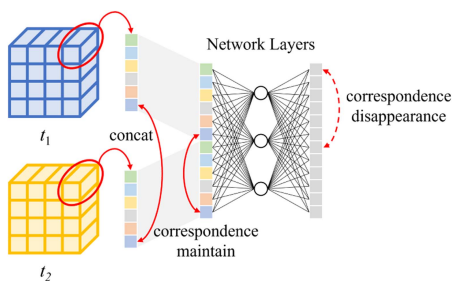

Fig. 2.   Correspondence disappearance in feature concatenation.

inferior to the single-branch networks, as shown in Fig. 1(e) and (f).

We argue that this is due to the initial dual-branch networks' inadequate utilization of the bitemporal spatial relationships during the processes of bitemporal feature extraction and fusion. During the feature extraction, the initial dual-branch networks independently extract features in each branch, which lack cross-branch interaction. Single-branch networks can achieve better results because they concatenate bitemporal inputs at the beginning. The network does not have the concept of bitemporal phases and can perceive information across them. Accordingly, researchers have designed several modules for cross-branch feature interaction. However, most of them concatenate bitemporal features and acquire features through network layers. As shown in Fig. 2, there should be correlation between features at the same location at different times, and after concatenating them, the next network layers have no spatial awareness in the dimension of depth, and will treat bitemporal features as a new feature vector, and ignore the relationship between the different temporal features. Therefore, the correspondence relationships among the bitemporal features would be disrupted. As shown in Fig. 1(g) and (h), while there has been some improvement, there remains substantial room for progression. In the feature fusion stage, existing approaches can be broadly categorized as module-based methods and subtraction-based methods. Existing module-based methods have the concatenation mentioned in

Fig. 2, resulting in the corresponding disappearance. While subtraction-based methods would lead to the features of unchanged regions approaching zero, causing a significant loss of information. Moreover, current methods have not accounted for the variability of different changes. These turn out to be bottlenecks of CD.

In order to tackle the previously mentioned concern, we propose a cross-branch interaction module within the feature extraction phase, along with a bitemporal feature fusion module. First, we propose the perception and interaction module (PIM), which constructs a dual-branch communication mechanism that can perform cross-branch perception without losing bitemporal spatial relationships, thereby extracting more precise and robust features. Second, we propose a patch–mode joint fusion module. With this module, we prevent the correspondence disappearance in recent fusion modules and the information loss in subtracting methods. We slice each patch in the bitemporal feature data into smaller mini-patches, and generate different fusion results for different change modes. By considering different changing modes and bitemporal spatial relationships, we could obtain more expressive fusion features. Finally, we have further optimized the network's backbone and redesigned its architecture based on our previous research RDP-Net [20]. We believe that our work offers a novel approach to utilizing the bitemporal spatial relationships in multitemporal remote sensing data, thereby enhancing the information mining capability of such data.

The SRC-Net has achieved a better performance in the field of remote sensing CD. The major contributions of SRC-Net can be summarized as follows.

1) We propose a PIM that utilizes the bitemporal spatial relationships among bitemporal inputs. It establishes a cross-branch interaction mechanism during the feature extraction process, leading to the extraction of more precise and robust features.

2) We propose a patch–mode joint feature fusion module (PM-FFM) that prevents the information loss present in current methods. Meanwhile, it considers different change modes and concerns about bitemporal spatial relationships, thereby obtaining more expressive fusion features.

3) Our proposed SRC-Net has been evaluated on two public datasets, LEVIR-CD and WHU Building datasets. The experimental results indicate that the proposed modules utilize bitemporal spatial relationships effectively, and our network achieved state-of-the-art (SOTA) performance.

The rest of this article is organized as follows. Section II provides a review of the pertinent literature concerning CD networks, feature interaction, and feature fusion. Section III describes the CD method proposed in this article. Section IV presents a set of quantitative comparisons and analyses based on experimental results. Section V presents discussion. Finally, Section VI concludes this article.

II. RELATED WORK

A. CD Networks

Neural networks have recently demonstrated promising outcomes in various geoscience applications [21], including scene

classification [3] and water quality estimates [22]. NN has also displayed impressive capabilities in CD [23], [24], [25]. Daudt et al. [26] utilized U-Net as a foundation and introduced early fusion (EF) and Siamese (Siam). They designed and proposed FC-EF, FC-Siam-conc, and FC-Siam-diff [19], which are widely acknowledged as the cornerstone of deep learning-based CD. Peng et al. [27] and Fang et al. [28] proposed UNet++_MSOF and SNUNet-CD based on FC-Siam-conc, replacing U-Net with U-Net++ [29] and merging multiple side outputs of U-Net++. Papadomanolaki et al. [30] presented an architecture similar to U-Net, referred to as L-UNet, which incorporates fully convolutional LSTM blocks at each encoding level to capture the temporal relationship of spatial feature representations. Chen and Shi [31] introduced a spatial–temporal attention mechanism to improve the discriminative capacity of deep features extracted from bitemporal images and proposed a late-fusion method STANet. As another attention-based method, Chen et al. [32] proposed a bitemporal image transformer (BIT) to accurately and efficiently capture contextual information within the spatial–temporal domain. Zhang et al. [33] introduced DSIFN, where they designed a shared deep feature extraction network as the encoder and a difference discrimination network as the decoder, demonstrating impressive performance. Chen et al. [20] proposed a region detail preserving network RDP-Net, avoiding information loss and improving the network's attention to details. Aiming at the loss of the spatial features of buildings caused by the multiple successive downsampling operations and the problem of incomplete buildings and the blurred edges caused by the complex scenes, Wang et al. [34], [35] proposed HDANet and W-Net. Zhou et al. [36] proposed a context aggregation network to mine interimage context over all training images for further enhancing intraimage context. Hang et al. [37] propose an ambiguity-aware network to address the difficulties in dealing with ambiguous regions, where pseudochanges happen or real changes are corrupted. Existing networks have often overlooked the bitemporal spatial relationships among bitemporal remote sensing data, which is the central motivation of our work.

B. Feature Interaction

There are two main ideas for dual-branch feature interaction. One utilizes a module to acquire dual-branch outputs from dual-branch inputs through neural layers, actualizing interaction among the branches via structure design. The other directly exchanges along spatial or feature dimensions of dual-branch feature maps to actualize dual-branch interaction.

The first idea has found considerable application in domains, such as multimodal correlation tasks [38], [39], registration [40], matching [41]. Lu et al. [42] introduced co-attention to aggregate different features from different branches and then distribute them as attention maps, respectively. Yu et al. [43] proposed SiamAttn to compute deformable self-attention and cross-attention from dual branches for object tracking. Wu et al. [40] proposed FIRE-Net, which aims to investigate the interaction of features within source and target point clouds across various levels.

Chen et al. [44] proposed MixFormer to achieve efficient feature interaction among windows and dimensions. Some recent networks, such as STANet, BIT, and BAN [45], already employ similar modules. These methods derive output results through computations across multiple network layers, which often require a considerable amount of computation. Meanwhile, they would result in the disappearance of the correspondence between bitemporal features, as shown in Fig. 2.

The other idea is directly exchanging along spatial or feature dimensions of dual-branch feature maps. CEN proposed by Wang et al. [46] is a representative, parameter-free multimodal fusion framework. This framework facilitates the dynamic exchange of channels among subnetworks of different modalities. The process of channel exchange is self-guided by the importance of individual channels. Fang et al. [47] proposed MetaChanger, directly exchanging features alternately in the spatial and channel dimensions. Such methods directly exchange dual-branch feature maps along spatial or feature dimensions, which would break the intrinsic relationships within the feature maps, and have not yet achieved widespread adoption.

In this article, we try to combine these two ideas, reducing the computation to an acceptable size without breaking the relationships within the original feature maps. Therefore, we propose a PIM, which calculates the credibility of the dual-branch feature maps separately and performs the interactions between branches based on the credibility.

C. Feature Fusion

Feature fusion modules are widely used in various deep learning applications, such as classification, semantic segmentation [29], [48], CD [19], [20], [28]. Existing modules are primarily for the fusion of multiscale and multimodal features [49], [50], [51], [52], [53]. They concatenate multiple feature matrices and then generate fusion results through the network layers, or vice versa. These modules neglect to consider the spatial relationships among bitemporal data. Mohammadian and Ghaderi [54] used the Key, Query, and Value matrices from the self-attention mechanism within Transformer to fuse bitemporal features and proposed SiamxFormer. The fusion result is obtained by a temporal transformer. However, the performance of SiamxFormer is not good enough.

In addition, within CD, there exists a simple feature fusion method that directly subtracts bitemporal data [32], [49]. This straightforward method has also achieved good fusion results. However, it causes the features of unchanged areas to approach zero, which leads to information loss and proves detrimental to the subsequent learning processes.

In this article, we introduce a novel feature fusion module that leverages the bitemporal spatial relationships of bitemporal remote sensing data, while avoiding information loss and taking into account the change modes.

III. METHODOLOGY

In this section, we first introduce the PIM, perceiving the spatial relationships among bitemporal inputs. Then, we propose
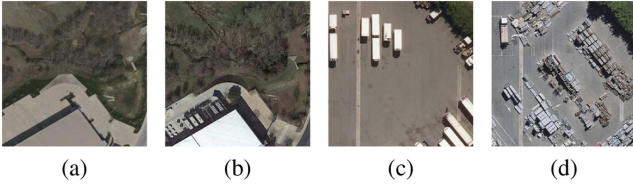
Fig. 3. Original input bitemporal remote sensing images. (a) and (b) LEVIR-CD dataset. (c) and (d) WHU dataset.

the patch–mode joint fusion module to avoid information loss and improve the expressive capability of the fusion features. At last, we present the architecture of the proposed SRC-Net.

A. Perception and Interaction Module

CD is the analysis of changes in land cover information, facing various challenges and situations. At times, factors, such as weather conditions and lighting, can impact the network's capacity to extract features and analyze the land cover information. In this context, the utilization of bitemporal comparisons can greatly facilitate the recognition of land cover information. As the RGB images shown in Fig. 3, the factory buildings at the bottom of Fig. 3(a) are similar to the nearby roads, making them susceptible to being mistaken for open space rather than industrial structures. The trucks situated in the top left corner of Fig. 3(c) are similar to rest areas or service stations, which are usually found on the roadside. The perception and interaction could help avoid misinterpreting the factory buildings as open space or the trucks as service stations. This example shows the importance of perception and interaction between branches in CD.

Therefore, we proposed the PIM. For the bitemporal feature maps extracted by the Siamese network, we define their credibility matrices separately, which refer to the probability that the feature map extracted from a single temporal instance is reliable. When the feature from a single temporal instance is deemed reliable, no adjustments would be made. However, if the feature is deemed unreliable, we would employ the feature from the other temporal instance to adjust it. To be more precise, if the other feature is deemed reliable, we would utilize it as the output; otherwise, we would utilize the average of the bitemporal features as the output. Drawing inspiration from the self-attention mechanism within Transformer [12], wherein the original input computes the Key, Query, and Value matrices directly via neurons, we similarly process bitemporal feature maps via neurons to acquire their credibility matrices separately. In practice, the credibility matrices can be viewed as the probability that the extracted features are correct, by which we can calculate the mathematical expectation of the correct features, and thus obtain the feature interaction results.

Specifically, when the inputs are two temporal instances, t_1 and t_2, we first calculate their credibilities separately, denoted as P_1 and P_2. Let t_1' and t_2' represent the output of this module. When t_1 is deemed reliable, $t_1' = t_1$. When t_1 is deemed unreliable and t_2 is deemed reliable, $t_1' = t_2$. When both t_1 and t_2 are deemed unreliable, $t_1' = \frac{t_1+t_2}{2}$. Let T_1 represent the event

where t_1 is deemed reliable, and $\overline{T_1}$ represent the event where t_1 is deemed unreliable. The same applies to T_2 and $\overline{T_2}$. The mathematical expectation of t_1' can be formulated as

$$
\begin{aligned}
E\left(t_1'\right) &= \sum_i \left(t_{1'(i)} P_{(i)}\right) \\
&= E\left(t_1' \mid T_1\right) P\left(T_1\right) + E\left(t_1' \mid \overline{T_1}, T_2\right) P\left(\overline{T_1}, T_2\right) \\
&\quad + E\left(t_1' \mid \overline{T_1}, \overline{T_2}\right) P\left(\overline{T_1}, \overline{T_2}\right) \\
&= t_1 P_1 + t_2\left(1 - P_1\right) P_2 + \frac{t_1+t_2}{2}\left(1 - P_1\right)\left(1 - P_2\right)
\end{aligned}
$$

(1)

where $i \in \{1, 2, 3\}$ denotes three potential situations: T_1, $\overline{T_1}T_2$, and $\overline{T_1T_2}$, $E(t_1' \mid T_1)$ denote the mathematical expectation of t_1' when t_1 is deemed reliable, and $P(T_1)$ denotes the possibility of T_1. The same applies to $E(t_1' \mid \overline{T_1}, T_2)$, $E(t_1' \mid \overline{T_1}, \overline{T_2})$, $P(\overline{T_1}, T_2)$, and $P(\overline{T_1}, \overline{T_2})$.

Above, we consider t_1 and t_2 as singular variables. However, in reality, t_1 and t_2 represent matrices as feature maps: $t_1, t_2 \in \mathbb{R}^{B \times C \times H \times W}$, where B, C, H, and W represent the batch size, channels, height, and width of the feature map. The credibilities of the feature maps $\{P_1, P_2\}$ are also two matrices: $P_1, P_2 \in \mathbb{R}^{B \times C \times H \times W}$. The multiplication in (1) employs the Hadamard product $\odot$ [55]. So, the mathematical expectation of t_1' is formulated as

$$
\begin{aligned}
E\left(t_1'\right) &= t_1 \odot P_1 + t_2 \odot \left(1 - P_1\right) \odot P_2 \\
&\quad + \frac{t_1+t_2}{2} \odot \left(1 - P_1\right) \odot \left(1 - P_2\right).
\end{aligned}
$$

(2)

Similarly, the mathematical expectation of t_2' can be formulated as

$$
\begin{aligned}
E\left(t_2'\right) &= t_2 \odot P_2 + t_1 \odot \left(1 - P_2\right) \odot P_1 \\
&\quad + \frac{t_1+t_2}{2} \odot \left(1 - P_2\right) \odot \left(1 - P_1\right).
\end{aligned}
$$

(3)

We name this feature interaction as the PIM. The credibility matrices are inferred from the input instances by linear layers, which can be formulated as follows:

$$
P_i = \text{sigmoid}\left(\text{Linear}(C, C)(t_i)\right)
$$

(4)

where $\text{Linear}(C, C)(\cdot)$ refers to a linear layer with C as input and output dimensions. The function $\text{sigmoid}(\cdot)$ maps a real number to the interval of (0,1).

The structure of the PIM is shown in Fig. 4. Our module takes into account the bitemporal spatial relationships in remote sensing CD, enabling the Siamese network to engage in dual-branch communication during the process of feature extraction. This enhances the network's ability to extract features from bitemporal remote sensing images.

B. Patch–Mode Joint Feature Fusion Module

One crucial task within CD is the fusion of feature maps extracted by Siamese networks from different temporal instances. Existing methods commonly employ concatenation or subtraction of bitemporal features. As mentioned previously,
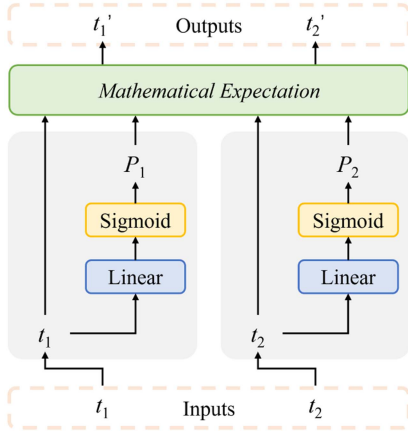
Fig. 4. Structure of the PIM.

existing methods inevitably result in information loss, and a lack of consideration for various situations of change. It is worth delving into bitemporal feature fusion to serve the subsequent prediction. The fusion module should take into account various situations of change and the bitemporal spatial relationships, thereby obtaining more expressive and robust fusion features.

Therefore, we propose the PM-FFM, inspired by the methods of handling time-varying signals in the field of signals and systems. In the realm of signal and system analysis, when examining a time-varying signal, it is common practice to extract the fundamental frequency or baseline of a time-varying signal first, which provides a macroscopic description of the signal's overall behavior. Derived from this, we can determine the mode of the signals, subsequently allowing for the selection of an optimal approach for signal processing. A similar idea should also be adopted for remote sensing CD. Different feature fusion heads should be used to fuse the bitemporal feature maps corresponding to different modes. The mean of bitemporal features can be regarded as the baseline, around which the features of preceding and subsequent temporal instances exhibit variations. Based on the baseline, we deduce the probabilities of various change modes, and obtain fusion features corresponding to these distinct modes through the specialized fusion heads. Subsequently, the final fusion result is ascertained based on mathematical expectation. Specifically, the structure of the patch–mode joint feature fusion module is shown in Fig. 5.

1) Bitemporal Subfeatures: We partition the input image into patches for learning at the beginning, so the size of the feature map extracted by the Siamese network can be equivalently represented as $d \times h \times w$, where h and w denote the height and width of the input feature map, respectively, and d denotes the depth. The size of each patch is $d \times 1 \times 1$. Considering that the modes within each patch may also vary, each patch is further subdivided into k mini-patches. The size of the mini-patches is $\frac{d}{k} \times 1 \times 1$. This allows for a more detailed determination of the modes within each patch. A mini-patch equivalently refers to a vector of dimensions $l \times 1$, where $l = d/k$. The mini-patches from the same position in the bitemporal feature maps are combined to form multiple sets of bitemporal subfeatures.

2) Mode Perception: A mode perception mechanism is constructed to determine the mode of the bitemporal subfeatures. As mentioned above, we consider the mean of the bitemporal subfeatures as the baseline and utilize a linear layer followed by a softmax layer to infer the probabilities of the different modes for each subfeature. The mode perception mechanism can be formulated as

$$M_i = \text{Softmax}\left(\text{Linear}(l, m)(\text{mean}(F_i))\right) \quad (5)$$

where $F_i = \{F_{i,1}, F_{i,2}\}$ represents the ith bitemporal subfeature, $F_{i,1}, F_{i,2} \in \mathbb{R}^{l \times 1}$ represent the ith subfeature of t_1 and t_2 feature map, $M_i \in \mathbb{R}^{m \times 1}$ represents the ith probability matrix of the different modes, and m represents the total number of modes. $\text{mean}(\cdot)$ refers to the operation of taking the mean value. $\text{Linear}(l, m)(\cdot)$ refers to a linear layer with l as input dimension and m as output dimension. $\text{Softmax}(\cdot)$ refers to the softmax function.

3) Multihead Feature Fusion: In correspondence with the various modes that exist for subfeatures, we design a multihead feature fusion mechanism. The inputs for different heads are obtained through linear combinations of the subfeature. Different combinations are processed through a group linear layer to obtain the fusion results of different modes. The number of groups in the group linear layer is equal to the number of modes, implying that each mode has its dedicated linear layer. This arrangement allows for generating unique feature fusion results for each mode. The combination of bitemporal subfeatures can be formulated as

$$C_{i,j} = \alpha_j F_{i,1} + \beta_j F_{i,2} \quad (6)$$

where α_j and β_j represent the jth coefficients for the linear combination of $F_{i,1}$ and $F_{i,2}$, respectively, and $C_{i,j} \in \mathbb{R}^{l \times 1}$ represents the jth linear combination, $j \in \{1, 2, \ldots, m\}$ of F_i. The multihead feature fusion results can be formulated as

$$H_{i,j} = \text{Linear}_{(j)}(l, l)(C_{i,j}) \quad (7)$$

where $\text{Linear}_{(j)}(l, l)(\cdot)$ refers to the jth linear layer with l as input and output dimension in the group linear layer. $H_{i,j} \in \mathbb{R}^{l \times 1}$ represents the feature fusion result of the jth head

$$H_i = \text{Concat}\left(H_{i,1}, H_{i,2}, \ldots, H_{i,m}\right)$$
$$= [H_{i,1}, H_{i,2}, \ldots, H_{i,m}] \quad (8)$$

where $\text{Concat}(\cdot)$ refers to concatenating matrices along the second dimension, and $H_i \in \mathbb{R}^{l \times m}$ represents the multihead feature fusion results.

Consequently, we obtain the probabilities of different modes for each subfeature, and obtain the fusion results by utilizing the feature fusion head specific to each mode. The final feature fusion result of a subfeature can be obtained through matrix multiplication, which can be formulated as

$$R_i = H_i \times M_i \quad (9)$$

where $R_i \in \mathbb{R}^{l \times 1}$ represents the final feature fusion result of the subfeature F_i. The fusion result of t_1 and t_2 feature maps is obtained by combining the results of every bitemporal subfeature.
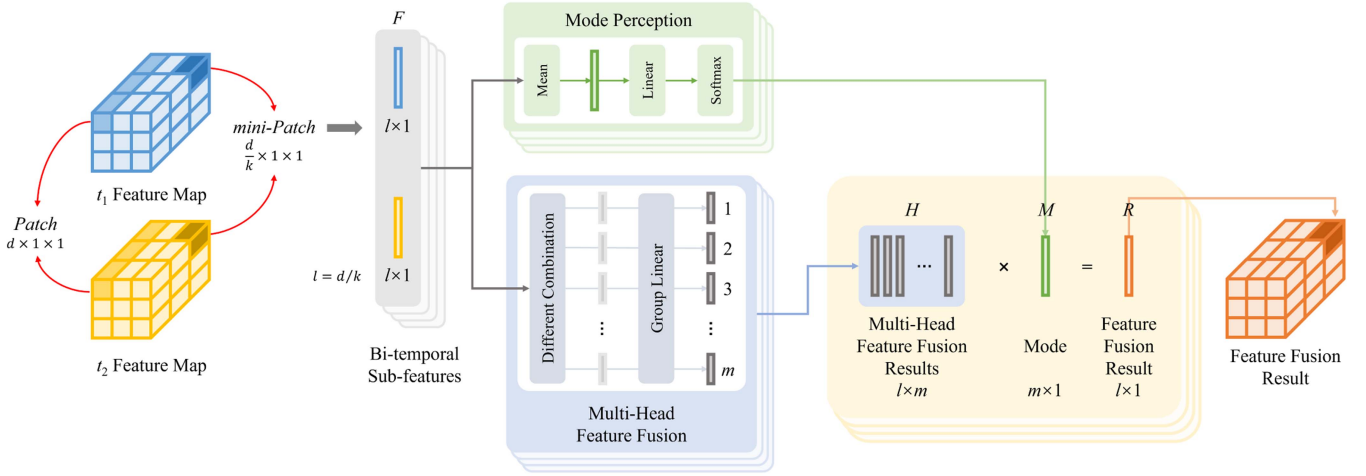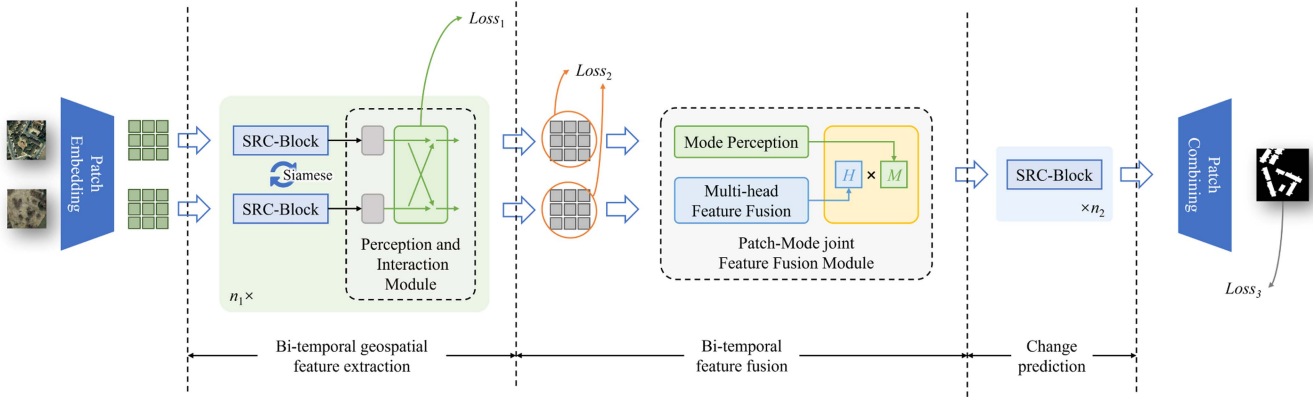
Fig. 5.   Structure of the PM-FFM.

Fig. 6.   Architecture of the proposed SRC-Net.

The proposed PM-FFM effectively avoids information loss, takes into account the different change modes, and utilizes the bitemporal spatial relationships of the input instances. In the experiment, the number of mini-patches in each patch k is set to 16, and the number of feature fusion heads m is also set to 4.

C. Network Architecture

The overall architecture of the proposed SRC-Net is shown in Fig. 6. The network backbone is similar to RDP-Net, but is enhanced by employing Siamese networks to extract features separately from bitemporal inputs, and we augment the supervision of network learning. Our network primarily consists of three stages: bitemporal geospatial feature extraction stage, bitemporal feature fusion stage, and change prediction stage, complemented by two additional layers: patch embedding layer and patch combining layer. In addition to supervising the final prediction, we have introduced supervision for the features extracted by the bitemporal geospatial feature extraction stage. The bitemporal features obtained in the feature extraction stage possess the capacity to characterize ground object information. By analyzing these features, we can identify the types of ground

objects at different times and initially determine the change areas. The comparison of the initial result with the ground truth ($Loss_2$) supervises the feature extraction stage.

Specifically, the patch embedding layer is employed to partition the input image into multiple patches based on their respective regions. In order to reduce the loss of information and maintain the intricate local details, the patch embedding layer with patch size $(p \times p)$ and input image size $(C \times H \times W)$ can be implemented as two convolutions ($\mathrm{GenPatch}(\cdot)$ and $\mathrm{GenPatch}_2(\cdot)$). The input channels of the first convolution $\mathrm{GenPatch}(\cdot)$ are c_{in}, output channels are $c/4$, kernel size is 4, and stride is 4. The input channels of the second convolution $\mathrm{GenPatch}_2(\cdot)$ are $c/4$, output channels are c, kernel size is $p/4$, and stride is $p/4$. The layer can be formulated as

$$X_{\mathrm{patch}} = \mathrm{GenPatch}_2\left(\mathrm{BN}(\mathrm{GenPatch}(X_{\mathrm{input}}))\right) \qquad (10)$$

where $\mathrm{BN}()$ represents Batch Normalization [56], X_{patch} represents the patch embedding feature maps, and X_{input} represents the input image.

The SRC-Block is used as the backbone of our network. As shown in Fig. 7, it consists of a depthwise convolutional layer followed by two pointwise convolutional layers. The depthwise
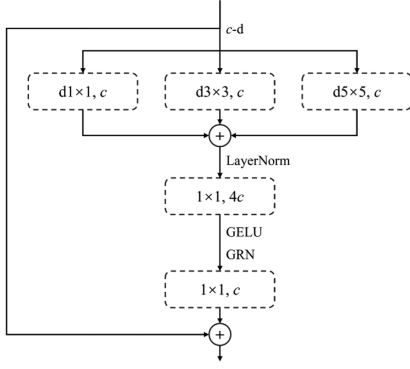
Fig. 7. SRC-Block.

layer consists of three parallelized group convolutional layers, which are used to explore the information between different patches. Each of the three group convolutional layers has c input and output channels, with c groups. The kernel sizes for these layers are 1, 3, and 5, respectively. The two pointwise convolutional layers are used to explore the information of each patch, where we explore the inverted bottleneck design. The kernel sizes for these convolutional layers are 1. The input channels and output channels for these layers are $(c, 4c)$, and $(4c, c)$, respectively. The residual connection serves to maintain a primary focus on the details inside a patch when the network explores the context information between patches. LayerNorm represents layer normalization [57], GELU represents Gaussian error linear unit [58], and GRN represents the global response normalization layer [59].

Bitemporal geospatial feature extraction stage: This stage employs Siamese networks to extract features, respectively, from bitemporal inputs. Each Siamese branch contains n_1 blocks, with each block consisting of a Siamese SRC-Block and a PIM. The SRC-Blocks on each branch of the Siamese networks share network parameters. These n_1 blocks can extract features from bitemporal inputs, harnessing the bitemporal spatial relationships and engaging in dual-branch communication.

Bitemporal feature fusion stage: This stage uses the PM-FFM, which takes into account the varying situations of changes, preserves bitemporal spatial relationships, and avoids information loss.

Change prediction stage: This stage contains n_2 SRC-Blocks to predict the change results from the fusion feature maps.

The patch combining layer is utilized to merge patches based on their respective regions in order to generate a pixel-level classification feature map. The region composition layer can be implemented as a convolution-transpose and a convolution. The input channels of the convolution-transpose $\mathrm{PatchUp}(\cdot)$ are c, output channels are 32, the kernel size is p, and stride is p. The input channels of the convolution $\mathrm{Final}(\cdot)$ are 32, output channels are out_ch, and the kernel size is 1. The layer can be formulated as

$$Y_{\text{output}} = \mathrm{Final}\left(\mathrm{GELU}\left(\mathrm{BN}(\mathrm{PatchUp}(Y_{\text{patch}}))\right)\right) \quad (11)$$

where Y_{output} represents the predicted results, and Y_{patch} represents the input patch embedding feature maps.

The previous paragraphs delineated the architecture of the different stages within the network. To facilitate a more robust learning process, we have implemented three sets of loss functions. These loss functions supervise the network's training together. The overall loss function can be defined as

$$\text{Loss} = \text{Loss}_1 + \text{Loss}_2 + \text{Loss}_3. \quad (12)$$

Loss_1 primarily imposes constraints on the training of the PIM in the bitemporal geospatial feature extraction stage, ensuring that the module can better retain pertinent information. Specifically, a singular temporal instance along with its noisy signal is fed into the module. The credibilities calculated for the noisy signal should be lower, and the credibilities of the original signal should be higher, resulting in two outputs that are closer to the original signal. We calculate the disparity between the module's output signal and the original data as the loss, which can be formulated as

$$\text{Loss}_1 = ||S_{\text{out}} - S_{\text{ori}}||_2 \quad (13)$$

where S_{out} represents the module's output signal and S_{ori} represents the original data.

Loss_2 and Loss_3 both describe the disparity between the prediction and the ground truth. Loss_2 primarily examines the feature extraction capability of the feature extraction network. Utilizing a convolution-transpose and a softmax layer, the likelihood of land cover types for each point can be inferred. Then, we can calculate the probability of change at each point by employing Bayes' formula, thereby computing the loss, which enables effective supervision of the feature extraction. Loss_3 is employed to calculate the disparity between the final prediction and the ground truth. In our network, we use a hybrid loss function as Loss_2 and Loss_3, which can be defined as $\mathcal{L}$

$$\mathcal{L} = \frac{1}{\sigma_1^2}\mathcal{L}_{\text{focal}} + \frac{1}{\sigma_2^2}\mathcal{L}_{\text{dice}} + \frac{1}{\sigma_3^2}\mathcal{L}_{\text{edge}} + \log \sigma_1\sigma_2\sigma_3 \quad (14)$$

where $\mathcal{L}_{\text{focal}}$ represents focal loss [60], $\mathcal{L}_{\text{dice}}$ represents dice loss [61], and $\mathcal{L}_{\text{edge}}$ represents edge loss [20]. σ_1, σ_2, and σ_3 represent the learnable relative weight of the losses $\mathcal{L}_{\text{focal}}$, $\mathcal{L}_{\text{dice}}$, and $\mathcal{L}_{\text{edge}}$, respectively, according to a principled approach to multitask deep learning proposed by Kendall et al. [62]. Focal loss and dice loss primarily address issues related to class imbalance. Edge loss is primarily concerned with capturing details, particularly in boundary regions. The formula of edge loss is

$$\mathcal{L}_{\text{edge}} = -w_{\text{edge}} \log(p_t) \quad (15)$$

where w_{edge} represents the edge weight of point and p_t represents the probability of correct classification.

IV. EXPERIMENTS AND ANALYSIS

A. Dataset

The experiment was conducted on two datasets named LEVIR-CD dataset [63] and WHU Building dataset [8], two of the most common datasets in remote sensing CD.

LEVIR-CD dataset consists of 637 visible light image pairs of 1024×1024 pixels, which are collected from Google Earth. The spatial resolution is 0.5m per pixel. We cut each image pair

TABLE I
RESULTS ON LEVIR-CD DATASET AND WHU BUILDING DATASET

Model	Params(M)	LEVIR-CD dataset					WHU Building dataset				
		Precision	Recall	F1	OA	IoU	Precision	Recall	F1	OA	IoU
FC-EF	1.35	0.7460	0.8826	0.8086	0.9825	0.6787	0.7842	0.7462	0.7647	0.9796	0.6191
FC-Siam-diff	1.35	0.8876	0.8789	0.8832	0.9902	0.7908	0.6740	0.8194	0.7993	0.9817	0.6656
FC-Siam-conc	1.55	0.8904	0.8899	0.8902	0.9908	0.8020	0.6106	0.8668	0.7165	0.9695	0.5582
STANet	16.97	**0.9338**	0.8658	0.8985	0.9910	0.8157	0.9190	0.8827	0.9005	0.9912	0.8190
BIT	11.94	0.9164	0.9170	<u>0.9169</u>	<u>0.9930</u>	<u>0.8466</u>	0.8959	<u>0.8965</u>	0.8962	0.9908	0.8119
L-UNet	8.45	<u>0.9318</u>	0.8864	0.9085	0.9919	0.8324	0.7900	0.8938	0.8387	0.9909	0.8144
DSIFN	50.71	0.8906	**0.9233**	0.9067	0.9920	0.8293	**0.9289**	0.8466	0.8859	0.9903	0.7951
SNUNet	12.03	0.9084	0.9110	0.9097	0.9924	0.8343	0.8826	0.8820	0.8823	0.9896	0.7894
RDP-Net	1.70	0.9148	0.8884	0.9012	0.9913	0.8125	0.9141	0.8893	<u>0.9015</u>	<u>0.9913</u>	0.8207
Changer	11.39	0.9249	0.9008	0.9127	0.9923	0.8394	0.9102	0.8855	0.8977	0.9909	0.8144
SiamixFormer	10.04	-	-	0.8947	-	-	-	-	0.8712	-	-
BAN	4.34	0.9290	0.8965	0.9124	0.9923	0.8389	0.9180	0.8860	0.9017	0.9913	<u>0.8210</u>
LightCDNet	2.82	0.9156	0.9000	0.9077	0.9918	0.8311	0.9083	0.8930	0.9006	0.9912	0.8191
SRC-Net	5.17	0.9263	<u>0.9172</u>	**0.9224**	**0.9935**	**0.8560**	<u>0.9257</u>	**0.9155**	**0.9206**	**0.9930**	**0.8528**

The first-best value is marked in bold, and the second-best value is underlined.

TABLE II
ABLATION EXPERIMENT ON LEVIR-CD DATASET

Method	PIM	PM-FFM	Precision	Recall	F1
SRC-Net$_\alpha$			0.9305	0.9025	0.9163
SRC-Net$_\beta$	✔		0.9249	0.9165	0.9206
SRC-Net$_\gamma$		✔	0.9265	0.9118	0.9191
SRC-Net	✔	✔	0.9263	0.9172	0.9224

TABLE III
EXPERIMENT WITH DIFFERENT k IN PM-FFM

k	2	4	8	**16**	32	64
Precision	0.9226	0.9257	0.9236	0.9263	0.9272	0.9267
Recall	0.9195	0.9185	0.9209	0.9172	0.9156	0.9164
F1	0.9210	0.9221	0.9222	0.9224	0.9214	0.9215

into 256×256 pixel patches without overlapping and ultimately obtained 7120 training sets and 1024 validation sets.

WHU Building dataset consists of two visible light aerial images of size $32\,507 \times 15\,354$, which are captured by NZ Aerial Mapping Ltd. The spatial resolution is 0.3 m per pixel. We cut each image pair into 256×256 pixel patches without overlapping and ultimately obtained 5141 training sets and 2293 validation sets.

B. Implementation Details

We implemented SRC-Net using the PyTorch framework. The number of PIM in the bitemporal geospatial feature extraction stage is set to 4, and the number of SRC-Blocks in the change prediction stage is also set to 4. c is set to 256. The learning rate is set to 2e-3 and decays by 0.8 every 20 epochs. In the training process, the batch size is set to 16, and AdamW [64] is applied as an optimizer. We conducted experiments on a single NVIDIA RTX3090 and trained for 300 epochs. The comparison methods also use all the same parameters.

We used three metrics, precision, recall, the F1 score, overall accuracy (OA), and IoU to quantitatively evaluate the performance of the CD models.

C. Comparison With SOTA Methods

We compared our method with FC-EF, FC-Siam-conc, FC-Siam-diff [19], RDP-Net [20], SNUNet [28], L-UNet [30], STANet [31], BIT [32], DSIFN [33], BAN [45], Changer [47], SiamixFormer [54], and LightCDNet [65]. These methods represent deep learning-based approaches in the field of CD. FC-EF, FC-Siam-conc, and FC-Siam-diff [19] are baseline models and they represent improvements of U-Net [48]. STANet [31] incorporates a spatial–temporal attention mechanism to improve the discrimination of deep features within bitemporal images. BIT [32] utilizes a customized CNN along with a pair of transformer encoders and decoders to effectively address the CD problem. L-UNet [30] represents a combination of convolutional and recurrent approaches, employing fully convolutional LSTM blocks for an end-to-end neural network. SNUNet [28], DSIFN [33], and LightCDNet [65] are three recently proposed methods, which are effective in obtaining competitive results. Changer [47] proposes a novel general architecture, MetaChanger, which integrates interaction layers within the process of feature extraction. SiamixFormer [54] leverages temporal transformers in feature fusion, enhancing the preservation of large receptive fields generated by transformer encoders. BAN [45] proposes a universal foundation model-based CD adaptation framework designed to extract foundational model knowledge for CD purposes.

Table I reports the comparisons of detection accuracy and the number of parameters. Our proposed SRC-Net shows better performance than other SOTA CD methods with only 5.17 M parameters. On the LEVIR-CD dataset, our SRC-Net achieved the highest F1 of 0.9224, surpassing other SOTA methods. The precision is 0.9263 and the recall is outstanding 0.9172, both of which are competitive, demonstrating a competitive performance. BIT secures the second position with an F1 of 0.9169, and 11.94 M parameters. SRC-Net has about 44% of BIT's parameters ($11.94\text{M} \times 43\% = 5.17\tilde{\ }\text{M}$). On the WHU Building dataset, our SRC-Net also achieved the highest F1 of 0.9224 and the highest recall of 0.9155, surpassing other SOTA methods. The precision is outstanding at 0.9257, further demonstrating a
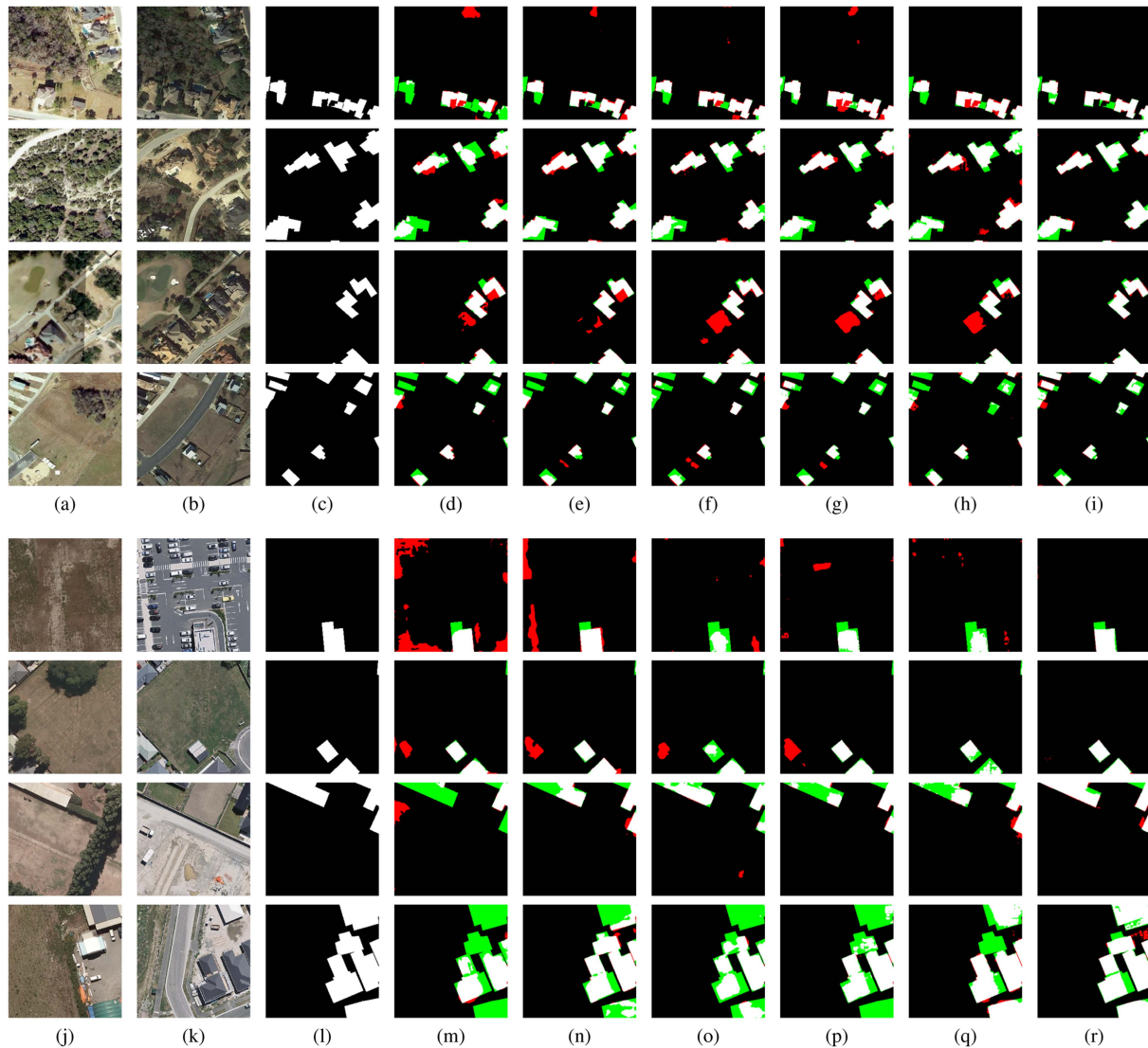
Fig. 8. (a)–(i) Results on the LEVIR-CD dataset. (j)–(r) Results on the WHU Building dataset. (a), (b), (j), and (k) Original images. (c) and (l) Ground truth. The results of (d) and (m) FC-Siam-diff, (e) and (n) BIT, (f) and (o) DSIFN, (g) and (p) SNUNet, (h) and (q) RDP-Net, and (i) and (r) our SRC-Net. The false positives and false negatives are indicated in red and green, respectively. Other colors represent true positives.

competitive performance. RDP-Net secures the second position, boasting an F1 of 0.9169. Fig. 8 shows some detection results from the validation set of the LEVIR-CD dataset and WHU Building dataset. The qualitative results provide a more intuitive comparison of the CD methods. We observe that most of the comparative methods perform poorly in complex scenes, such as the case of tree occlusion. In contrast, thanks to the more accurate features extracted by the PIM and the more expressive features fused by the PM-FFM, our SRC-Net has better performance.

In our experiments, particularly in certain scenarios, we have also observed significant improvements in our SRC-Net compared with other SOTA methods, as shown in Fig. 9. For scenarios where the optical properties of the road and the roofs of the surrounding buildings are similar, comparative methods perform poorly, and our SRC-Net still gives good results. However, we also note that there is still room for improvement of our method in this scenario. These experiments further prove the superiority

of our SRC-Net. We believe that these improvements are the result of our enhancements, which will be analyzed in detail later.

D. Ablation Study

To evaluate the PIM and PM-FFM, a couple of ablation experiments were conducted, as given in Table II. There are two variants of our network, in addition to the full network. They are as follows.

1) *SRC-Net$_\gamma$:* We remove PIM from SRC-Net.
2) *SRC-Net$_\beta$:* We remove the PM-FFM from SRC-Net, and just subtract bitemporal feature maps to obtain the new feature map.
3) *SRC-Net$_\alpha$:* We remove PIM from SRC-Net$_\beta$.

The experimental results exemplify the contributions of each improvement in our SRC-Net. The PIM improves our network
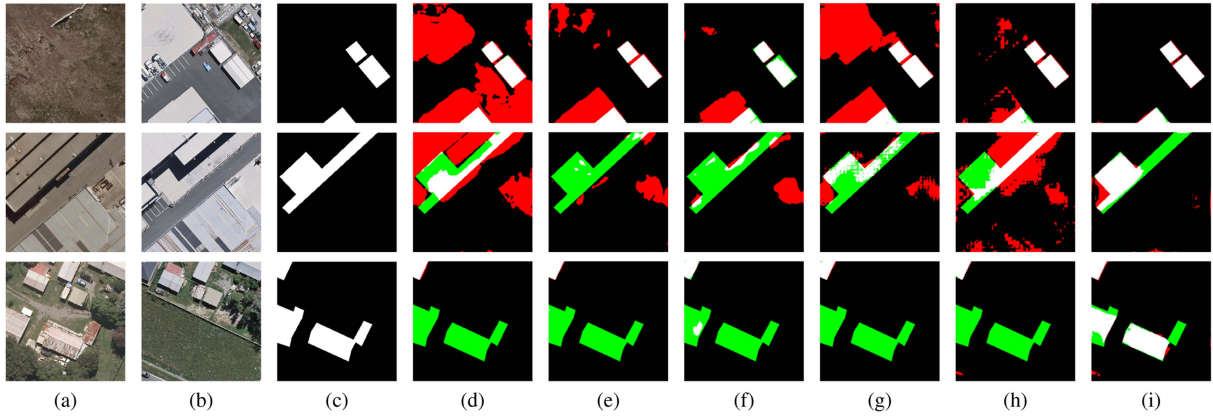
Fig. 9.   (a) and (b) Original images. (c) Ground truth. The result of (d) FC-Siam-diff, (e) BIT, (f) DSIFN, (g) SNUNet, (h) RDP-Net, and (i) our SRC-Net. The false positives and false negatives are indicated in red and green, respectively. Other colors represent true positives.
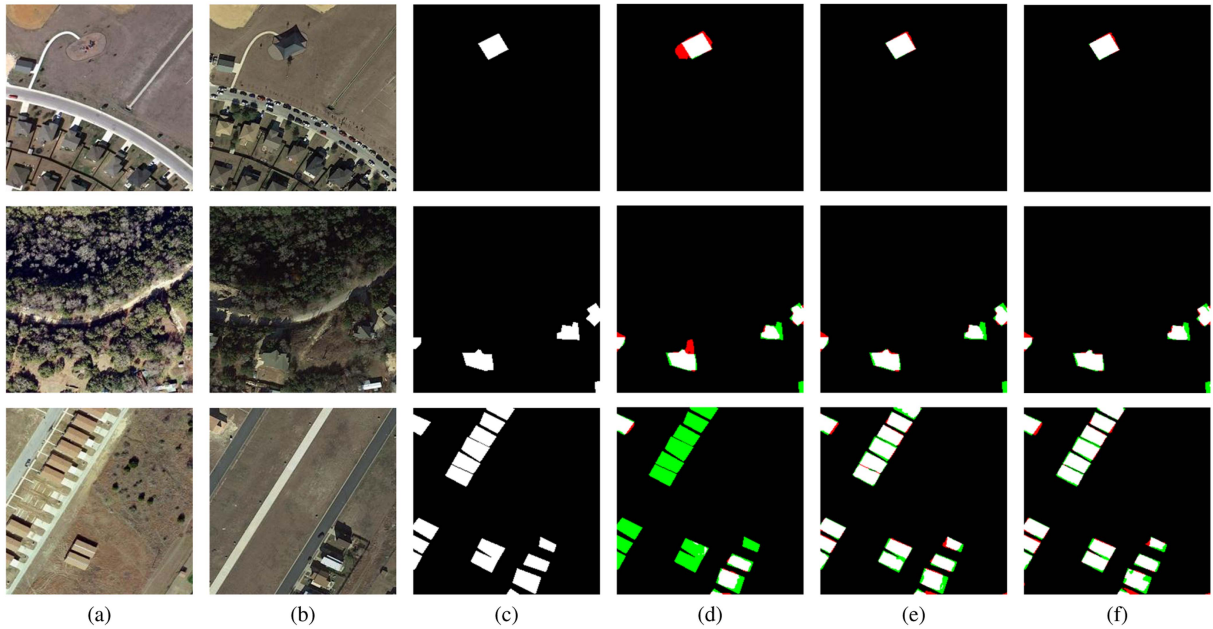

Fig. 10.   (a) and (b) Original images. (c) Ground truth. The result of (d) SRC-Net$_\alpha$, (e) SRC-Net$_\beta$, and (f) SRC-Net. The false positives and false negatives are indicated in red and green, respectively. Other colors represent true positives.
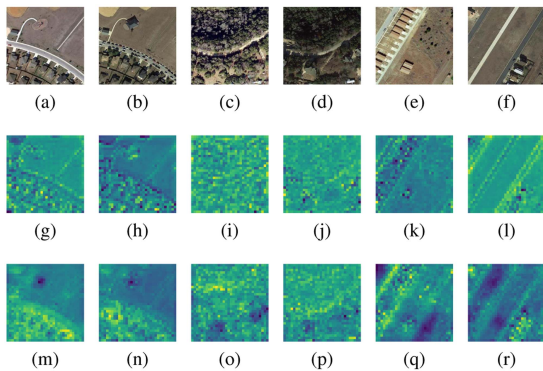

Fig. 11.   (a)–(f) Original images. The visualization results of (g)–(l) the output feature maps of the bitemporal geospatial feature extraction stage of SRC-Net$_\alpha$, and (m)–(r) SRC-Net$_\beta$.

by 1.40% recall and 0.43% F1. The PM-FFM improves our network by 0.93% recall, and 0.28% F1. Fig. 10 shows some detection results.

The experimental results suggest that our improvements can increase recall, thereby allowing the network to detect more change areas. In addition, the experiments demonstrate the efficiency of the SRC-Net architecture, enabling the attainment of performance levels comparable with SOTA standards even in the absence of the PIM and PM-FFM.

E. Experiment in PIM

The aforementioned experiments collectively demonstrate the remarkable superiority of our SRC-Net. By analyzing experimental results, we conclude that a significant factor lies in the better feature extraction capabilities during
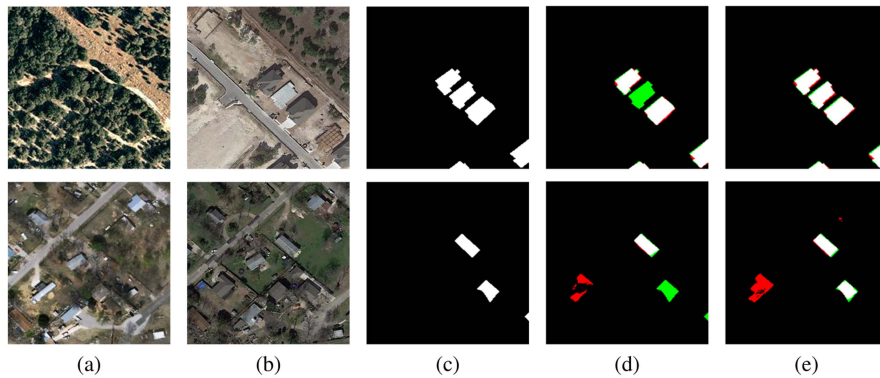
Fig. 12. (a) and (b) Original images. (c) Ground truth. The result of (d) SRC-Net$_\beta$ and (e) our SRC-Net. The false positives and false negatives are indicated in red and green, respectively. Other colors represent true positives.

the bitemporal geospatial feature extraction stage. The PIM facilitates dual-branch communication. These factors contribute to the network's extraction of more precise and robust features.

As shown in Figs. 9 and 10, SRC-Net exhibits enhanced discernment when it comes to buildings that are easily confused with their surrounding environments. We have visualized the bitemporal feature maps obtained by extracting the instances in Fig. 10 through SRC-Net$_\alpha$ and SRC-Net$_\beta$, and the results are shown in Fig. 11. From the results, it is evident that for regions with changes in buildings, the feature maps extracted by SRC-Net$_\beta$ exhibit more pronounced variations compared with SRC-Net$_\alpha$. Furthermore, the delineation of the area is more precise. These are highly advantageous for subsequent predictions of changed areas.

F. Analysis of PM-FFM

The fusion module is designed to prevent the information loss present in current methods. Meanwhile, it considers different change modes and concerns about bitemporal spatial relationships, thereby obtaining more expressive fusion features.

For the setting of the hyperparameter k in the module, six values from 2 to 64 are selected and tested on the LEVIR-CD dataset. The results are given in Table III. It can be seen that when k is larger than 4, the impact on the final performance will not be significant, and we set k to 16 based on the result.

In addition, we visualize the results of bitemporal feature extraction. As shown in Fig. 12, in the top row of images, the area surrounding the three houses in the middle remains as land before and after the change. However, the middle house differs from the other buildings, yet it might be misidentified as flat land when looking at a single image. If we observe the fusion features by direct subtraction, crucial features that can determine the type of this area may be lost. As a result, SRC-Net$_\beta$ is unable to identify this particular change. With the RS-CD feature fusion module, this change is well predicted. In the bottom row, this situation becomes even more apparent. The input images have significant forest cover, and if we simply subtract them, it becomes challenging for the subsequent network to perceive

the surrounding environment's characteristics. It may fail to recognize the potential noise that could have been introduced in the previous network.

V. DISCUSSION

Most existing CD networks lack structure design for perceiving bitemporal spatial relationships. Therefore, existing methods are insufficient in the utilization of bitemporal spatial relationships, which is crucial for CD. The SOTA methods compared in Table I can be broadly categorized into three groups: single-branch networks (FC-EF, L-UNet, SNUNet, and RDP-Net), dual-branch networks without cross-branch interaction (FC-Siam-conc, FC-Siam-diff, STANet, SiamixFormer, and BAN), and dual-branch networks with cross-branch interaction (BIT, DSIFN, Changer, LightCDNet, and SRC-Net). From the results, it can be seen that dual-branch networks without cross-branch interaction achieve generally poorer results than single-branch networks, whereas dual-branch networks with cross-branch interaction achieve better results. The results further validate the importance of bitemporal spatial relationship. The two modules proposed in this article significantly enhance the network's ability to perceive spatial correlation during the feature extraction and feature fusion stages.

However, the current SRC-Net still has some limitations. The SRC-Block can be reduced in the amount of computation and the number of parameters. In addition, the PIM relies on the learned credibility matrices for interaction, ignoring that the land cover may have changed between the two time phases. These issues will be the focus of our future research.

VI. CONCLUSION

In this article, we propose SRC-Net, an effective method concerning bitemporal spatial relationships for remote sensing CD. It contains the PIM and the PM-FFM. These modules effectively utilize the bitemporal spatial relationships in the input bitemporal data. The PIM establishes a cross-branch perception mechanism during the feature extraction process with the bitemporal spatial relationships among bitemporal inputs, leading to extract more precise and robust features. The PM-FFM

considers different change modes, and concerns bitemporal spatial relationships, and prevents information loss in current methods. With this module, we can obtain more expressive fusion features. We have corroborated the effectiveness of our network through comprehensive experiments. The proposed SRC-Net achieves the SOTA empirical performance with 5.17 M parameters. Experiments were conducted on the LEVIR-CD and WHU Building datasets, with the results indicating that the SRC-Net model has the potential to notably improve the accuracy of remote sensing CD. However, we also note that the computational overhead of the current backbone of SRC-Net remains a limiting factor for deployment on edge devices. The research in this article is based on bitemporal remote sensing images, but the two modules proposed can be also applied in multitemporal CD tasks, which will be the direction of our subsequent research.

REFERENCES

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] X. Li, F. Pu, R. Yang, R. Gui, and X. Xu, "AMN: Attention metric network for one-shot remote sensing image scene classification," *Remote Sens.*, vol. 12, no. 24, 2020, Art. no. 4046.

[3] R. Yang, F. Pu, Z. Xu, C. Ding, and X. Xu, "DA2Net: Distraction-attention-driven adversarial network for robust remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 21, 2021, Art. no. 8012705.

[4] M. Reba and K. C. Seto, "A systematic review and assessment of algorithms to detect, characterize, and monitor urban land change," *Remote Sens. Environ.*, vol. 242, 2020, Art. no. 111739.

[5] S. Sippel, N. Meinshausen, E. M. Fischer, E. Székely, and R. Knutti, "Climate change now detectable from any single day of weather at global scale," *Nature Climate Change*, vol. 10, no. 1, pp. 35–41, 2020.

[6] Y. Sun, X. Zhang, Y. Ding, D. Chen, D. Qin, and P. Zhai, "Understanding human influence on climate change in China," *Nat. Sci. Rev.*, vol. 9, no. 3, 2022, Art. no. nwab113.

[7] K. M. Sim, "Grid commerce, market-driven G-negotiation, and grid resource management," *IEEE Trans. Syst. Man Cybern. Part B.*, vol. 36, no. 6, pp. 1381–1394, Dec. 2006.

[8] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multi-source 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.

[9] 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.

[10] 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.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE Proc. IRE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[12] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5999–6009, 2017.

[13] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogrammetry Remote Sens.*, vol. 80, pp. 91–106, 2013.

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

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

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

[17] S. Saha, Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "Unsupervised deep transfer learning-based change detection for HR multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 856–860, May 2021.

[18] S. Saha, F. Bovolo, and L. Bruzzone, "Building change detection in VHR SAR images via unsupervised deep transcoding," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1917–1929, Mar. 2021.

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

[20] H. Chen, F. Pu, R. Yang, R. Tang, and X. Xu, "RDP-Net: Region detail preserving network for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 6, 2022, Art. no. 5635010.

[21] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher, "A critical synthesis of remotely sensed optical image change detection techniques," *Remote Sens. Environ.*, vol. 160, pp. 1–14, 2015.

[22] Y. Hu, F. Pu, C. Kong, R. Yang, H. Chen, and X. Xu, "A2DWQPE: Adaptive and automated data-driven water quality parameter estimation," *J. Hydrol.*, vol. 626, pp. 1–13, 2023, Art. no. 130363.

[23] 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.

[24] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12084–12098, Nov. 2022.

[25] Z. Zhu, "Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 370–384, 2017.

[26] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[27] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1382.

[28] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 17, 2021, Art. no. 8007805.

[29] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany:Springer, 2018, pp. 3–11.

[30] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.

[31] 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.

[32] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 20, 2021, Art. no. 5607514.

[33] 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.

[34] X. Wang et al., "A high-resolution feature difference attention network for the application of building change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102950.

[35] X. Wang et al., "Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, 2023, Art. no. 103456.

[36] F. Zhou, C. Xu, R. Hang, R. Zhang, and Q. Liu, "Mining joint intraimage and interimage context for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 12, 2023, Art. no. 4403712.

[37] R. Hang, S. Xu, P. Yuan, and Q. Liu, "AANet: An ambiguity-aware network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 29, 2024, Art. no. 5612911.

[38] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6087–6096.

[39] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 561–577.

[40] B. Wu, J. Ma, G. Chen, and P. An, "Feature interactive representation for point cloud registration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5530–5539.

[41] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10941–10950.

[42] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 289–297, 2016.

[43] Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, "Deformable Siamese attention networks for visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6728–6737.

[44] Q. Chen et al., "MixFormer: Mixing features across windows and dimensions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5249–5259.

[45] K. Li, X. Cao, and D. Meng, "A new learning paradigm for foundation model-based remote-sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 16, 2024, Art. no. 5610112.

[46] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 4835–4845, 2020.

[47] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 18, 2023, Art. no. 5610111.

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

[49] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3560–3569.

[50] G. Feng, Z. Hu, L. Zhang, and H. Lu, "Encoder fusion network with co-attention embedding for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15506–15515.

[51] W. G. C. Bandara and V. M. Patel, "HyperTransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1767–1777.

[52] H. Zhou, L. Qi, H. Huang, X. Yang, Z. Wan, and X. Wen, "CANet: Co-attention network for RGB-D semantic segmentation," *Pattern Recognit.*, vol. 124, 2022, Art. no. 108468.

[53] Z. Huang, Y. Wei, X. Wang, W. Liu, T. S. Huang, and H. Shi, "AlignSeg: Feature-aligned segmentation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 550–557, Jan. 2022.

[54] A. Mohammadian and F. Ghaderi, "SiamixFormer: A fully-transformer Siamese network with temporal Fusion for accurate building detection and change detection in bi-temporal remote sensing images," *Int. J. Remote Sens.*, vol. 44, no. 12, pp. 3660–3678, 2023.

[55] R. A. Horn, "The Hadamard product," in *Proc. Symp. Appl. Math*, 1990, pp. 87–169.

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[57] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[58] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[59] S. Woo et al., "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16133–16142.

[60] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[61] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[62] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.

[63] 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. [Online]. Available: https://www.mdpi.com/2072-4292/12/10/1662

[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018.

[65] Y. Xing, J. Jiang, J. Xiang, E. Yan, Y. Song, and D. Mo, "LightCDNet: Lightweight change detection network based on VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Aug. 11, 2023, Art. no. 2504105.

Hongjia Chen received the M.S. degree in information and communication engineering in 2022 from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in signal and information processing.

His research interests include machine learning and signal processing.

Xin Xu received the B.S. degree in electronic engineering, the M.S. degree in signal and information processing, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1989, 1996, and 2003, respectively.

He is currently a Professor with Electronic Information School, Wuhan University. His research interests include image understanding and statistical signal processing with applications to remote sensing.

Fangling Pu (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005.

In 2007, she was a Research Assistant with the Institute for Theoretical Information Technology, RWTH Aachen University. Since 2000, she has been an Associate Professor with Electronic Information School, Wuhan University. Her research interests include system and networking, signal and information processing, and data mining.