Siamese Biattention Pooling Network for Change Detection in Remote Sensing

Hengzhi Chen, Kun Hu, Patrick Filippi, Wei Xiang, *Senior Member, IEEE*, Thomas Bishop, and Zhiyong Wang, *Member, IEEE*

***Abstract*—Change detection (CD) in remote sensing aims to identify variations in image pairs captured at the same location but different times. While recent deep learning approaches, particularly those incorporating attention mechanisms, have achieved encouraging results on this task, they often fall short of comprehensively exploiting the change relevant patterns that are present in paired images. In this study, we propose a novel deep learning architecture, namely Siamese Bi-Attention Pooling Network (SBA-PN), to emphasize broad-scale change patterns by exploiting both intraimage and interimage contexts. The overall structure of SBA-PN aligns with the U-Net based encoder-decoder paradigm. A Siamese Transformer-like encoder formulates paired multiscale feature maps. To effectively emphasize change relevant patterns, a spatial optimal pooling module is devised, replacing the conventional self-attention mechanism. A contrastive pixel-wise supervision scheme is designed for shallow encoder layers in pursuit of change-aware feature maps. Next, the decoder mirrors the multiscale design, which formulates difference maps using a novel biattention mechanism from paired feature maps. During the decoding phase, a channel deviation pooling module is devised to further emphasize salient change regions. Comprehensive experimental results demonstrate the effectiveness of the proposed method with the state-of-the-art performance on two commonly used benchmark datasets, Sun Yat-Sen University (SYSU)-CD and LEarning VIsion Remote sensing (LEVIR)-CD.**

***Index Terms*—Attention, change detection (CD), deep learning (DL), remote sensing.**

I. Introduction

OUR dynamic Earth undergoes continuous changes, due to both natural occurrences and human activities. Therefore, monitoring the health of the Earth and promoting sustainable human development necessitate better understanding of these changes. Change detection (CD) in remote sensing serves this purpose, which identifies disparities in land surfaces by comparing two remote sensing images acquired from the same geographical area, but at different points of time. Change detection is a cornerstone in remote sensing image analysis and plays an important role in extensive applications, such as monitoring urban expansion [1], disaster management [2], and deforestation control [3]. As satellite technology continues to advance, an array of sensors with increasing functionalities such as active or passive, optical or microwave, and high or low resolution, have been launched into space [4]. The wealth of high-quality remote sensing data facilitates various studies, yet it also brings forth new challenges that need to be addressed for change detection.

Conventional change detection methods rely primarily on spectral information extracted from remote sensing images. Existing methods, such as image differencing [7], [8], regression analysis [9], change vector analysis (CVA) [10], [11] and principal component analysis (PCA) [12], although effective, involve extensive image preprocessing, optimal thresholding, and the integration of handcrafted features. These requirements often limit their robustness and efficiency in real-world applications. The introduction of deep learning (DL) to change detection addresses these challenges. The core advantage of DL lies in its capability to learn intricate patterns or representations without the necessity of manual feature engineering. By learning directly from labeled samples, DL offers a data-centric approach to modeling, eliminating the dependency on explicit handcrafted features and subsequently enhancing the potential accuracy and adaptability of change detection models.

Recent DL methods for change detection have shifted focus to a larger receptive field, recognizing that changes, particularly in remote sensing imagery, are context-dependent. A neural network with an expanded receptive field can better assimilate information from a wider spatial context. This improves its sensitivity to subtle and contextually significant changes. Techniques such as stacking additional convolutional layers [13] have been employed to achieve this. Moreover, the integration of attention mechanisms [14], [15], [16], [17], [18], [19], [20], [21], [22] has marked a significant advancement in this domain. Drawing parallels to human perception, where our focus naturally gravitates towards salient features in a scene [23], attention mechanisms in DL models enable selective emphasis on critical information. This translates to neural networks that prioritize relevant change characteristics over extraneous data. By automating feature map weighting through these mechanisms, networks can pinpoint pivotal change signals, leading to an improvement in detection precision.

Existing attention-based methods in change detection can be broadly classified into three categories: 1) spatial-wise

Manuscript received 3 December 2023; revised 30 January 2024; accepted 19 February 2024. Date of publication 8 March 2024; date of current version 2 April 2024. *(Corresponding author: Kun Hu.)*

Hengzhi Chen, Kun Hu, and Zhiyong Wang are with the School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: hche6958@uni.sydney.edu.au; hukun_sdu@hotmail.com; zhiyong.wang@sydney.edu.au).

Patrick Filippi and Thomas Bishop are with the School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: patrick.filippi@sydney.edu.au; thomas.bishop@sydney.edu.au).

Wei Xiang is with the School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC 3086, Australia (e-mail: w.xiang@latrobe.edu.au).

Digital Object Identifier 10.1109/JSTARS.2024.3373753

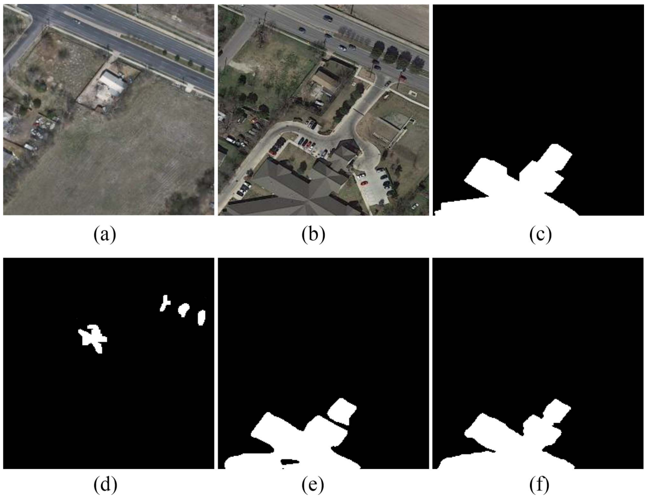

Fig. 1. Comparative visualization of change detection techniques. (a) Original image captured at time t_1. (b) Subsequent image captured at time t_2. (c) The ground truth change map for reference. (d) Change map derived from a convolutional deep learning method (2018) [5]. (e) Change map produced using a Transformer-based deep learning method (2022) [6]. (f) Change map produced using SBA-PN (ours).

attention [16], [17]; 2) channel-wise attention [14], [15]; and 3) mixed attention [18], [19]. Spatial-wise attentions allow neural networks to emphasize specific regions within an image, targeting pertinent or anomalous areas. Channel-wise attentions, on the other hand, concentrate on distinctive channels, accentuating those of particular relevance to the context. Adeptly leveraging the benefits of these two approaches, mixed attention integrates spatial-wise attention to pinpoint critical spatial anomalies and utilizes channel-wise attention to emphasize pertinent feature details.

While these attention-based methods provide significant improvements in change detection, their effectiveness is inherently limited by their local scope of operation. Theoretically, these methods operate within a constrained contextual window, dictated by the fixed kernel size of convolution operations, making it challenging to encapsulate the global context necessary for comprehensive change detection. This local focus leads to a substantial theoretical limitation: The inability to capture the long-range, interpixel dependencies and contextual nuances critical in accurately representing the complex, dynamic interactions in remote sensing imagery.

Several recent studies [20], [21] have reported encouraging results using Transformer-based models to capture the semantic relationships between pixels. The original design of Vision Transformers (ViTs) primarily focuses on handling standalone image inputs. However, in applications like change detection, where it is necessary to analyze paired image features, this approach encounters limitations. Methods such as feature differencing or concatenation are commonly used for this purpose, however they can result in the loss of fine-grained information due to their inherent coarseness. This is illustrated in Fig. 1(e), where the change map, generated by the method from [6], presents noticeable sparkling discontinuities, indicating the loss of details during the merging process.

Motivated by the aforementioned challenges, this article introduces the Siamese Bi-Attention Pooling Network (SBA-PN), a novel DL method designed to enhance the single-input self-attention mechanism. The SBA-PN facilitates interaction with feature pairs and incorporates feature differencing, enabling the comprehensive capture of long-range dependencies and fine-grained details. Importantly, it achieves this sophistication without incurring high computational costs by efficiently eliminating redundant information. The overall structure of the SBA-PN is a U-Net based encoder-decoder architecture, as illustrated in Fig. 2. First, a bitemporal image pair is fed into a Siamese Transformer-like encoder with a *spatial optimal pooling* mechanism instead of a conventional self-attention mechanism, which aims to effectively highlight the change detection hints in the resultant feature maps. Particularly, the encoder is aided with a multistage design to generate multiscaled paired semantic feature maps. This corresponds with the decoder design that progressively computes the distance maps between the feature maps from each encoding stage. For comprehensive decoding in each stage, a novel biattention mechanism is introduced to formulate a difference map based on the encoded feature map pair and the output from the previous decoding stage. To enhance the salient information during decoding, a standard deviation-based operation, *channel deviation pooling* is devised which is able to reduce the channel length and encourages the network to prioritize semantically rich channels. Finally, a prediction head (i.e., change detection head) is used to produce the final change map. In addition, compared to the conventional training pipeline for change detection, we introduce a novel contrastive pixel-wise supervision scheme for shallow encoder layers in pursuit of change-aware feature maps.

Overall, the key contributions of this article are as follows.

1) We propose a novel SBA-PN network, designed to exploit the paired relations in bitemporal images across both spatial and channel dimensions, utilizing straightforward yet efficient pooling strategies.
2) We devise a multistage encoder with a novel spatial optimal pooling (SOP)-Former architecture, uniquely introducing SOP to formulate spatial patterns instead of using a conventional self-attention mechanism. Our SOP mechanism able to achieve an elegantly simple balance between max and average pooling, enhancing the pooling operation's flexibility and context-awareness. Our decoder employs a biattention mechanism optimized for processing paired image features. In addition, we introduce channel deviation pooling (CDP), a standard deviation-based pooling strategy. CDP effectively reduces redundancy by selectively filtering out less informative features across the channel dimension.
3) A contrastive pixel-wise supervision schema is devised for shallow network layers, optimizing the extraction of change-aware feature maps.
4) Comprehensive experiments were conducted to demonstrate the effectiveness of our proposed method on two widely used datasets–LEVIR-CD and SYSU-CD with the state-of-the-art performance.

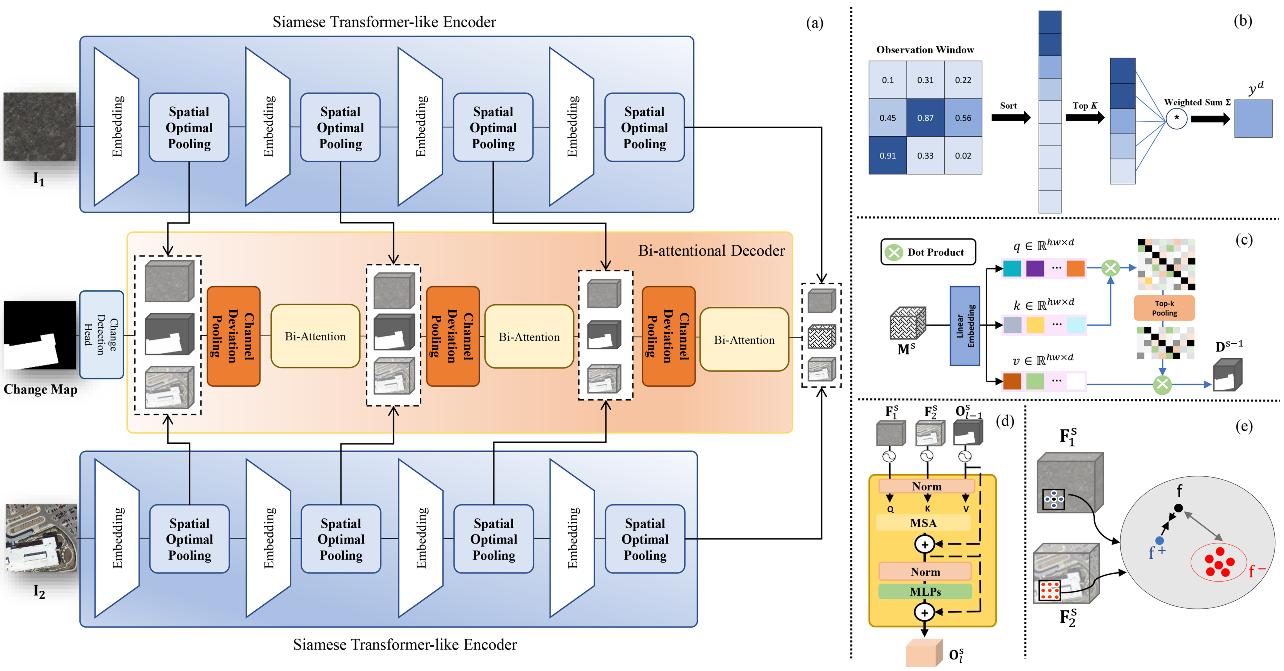

Fig. 2. Illustration of the proposed SBA-PN architecture. (a) SBA-PN follows a U-Net based encoder-decoder design, which consists of a Siamese Transformer-like encoder to formulate paired feature maps in a multiscale manner, and a biattention-based decoder corresponds to this multiscale design, which derives difference maps. (b) The proposed SOP. (c) The proposed CDP. (d) The proposed biattention mechanism. (e) The illustration of Contrastive Pixel-wise Supervision.

The rest of this article is organized as follows. Section II reviews the related work of traditional methods and DL methods. Section III provides the details of our proposed method. The experimental results are presented in Section IV with the discussions and analyses. Finally, Section V concludes this article.

II. Related Work

CD techniques have evolved from traditional methods to advanced DL-based approaches. In the early days of CD research, image differencing and CVA were commonly used. While these traditional methods were effective for change detection, they had limitations in handling complex spatial and temporal variations [3], [24]. The shift to DL models marked a significant advancement in CD, benefiting from hierarchical feature extraction and end-to-end learning paradigms. More recently, Transformer-based architectures have emerged in the CD domain. These models have shown promise in detecting intricate changes across varied datasets. This section reviews the development of CD across three major phases: 1) traditional CD methods; 2) DL CD methods; and 3) Transformer-based methods.

A. Traditional CD Methods

Image differencing [7], [8] was a foundational approach in CD, employing a co-registered multitemporal image pair to generate a residual image that highlighted changes. Changes were either discerned directly from pixel values or based on preprocessed metrics like vegetation indices. However, this method's need for manual thresholding and occasional inaccuracies in determining change magnitude were evident limitations. Regression analysis between bitemporal images [9] assumed a linear projection from the preevent to the postevent image, but identifying an accurate linear function remained a challenge for subtle bitemporal changes [25]. In contrast to singular band-based methods, which analyze satellite imagery one spectral band at a time, CVA examines multiple bands simultaneously [10], [11]. Using vectors to encapsulate pixel values across spectral bands, CVA computed change vectors by subtracting vectors from different observation points. The direction of these vectors indicated change type and their magnitude revealed change extent. Yet, CVA demanded consistent atmospheric conditions for the image pairs–a challenging prerequisite in many remote sensing contexts. Reducing redundancy in imagery, PCA was also integrated into CD [12], [26]. Here, image data was reconfigured into principal components based on variance, with the first two typically denoting unchanged regions. Moving from conventional techniques like PCA, the Bi-CCD method [27] represents a significant advancement, improving land cover monitoring accuracy with its bidirectional analysis, thereby reducing common errors seen in traditional CD approaches. Despite their advancements, traditional CD methods confronted challenges, from manual thresholding to linearity assumptions and reliance on handcrafted features. These challenges underscored their limitations in handling high-dimensional data and providing meaningful semantic insights.

B. Deep Learning CD Methods

DL has established itself as a groundbreaking advancement, addressing the complexities inherent in traditional CD methods. Distinct from their predecessors, DL models autonomously extract salient features from paired images, negating the need for elaborate handcrafted features. FCN [5] first proposed an end-to-end convolutional network designed for simultaneous feature extraction from paired images, leading to change maps derived from various feature differencing methods. Subsequent work [13] has incorporated advanced convolutional layers to produce feature difference maps across diverse scales, mitigating the presence of pseudochanges. In addition, MTCDN [28], which integrates optical and SAR imagery for change detection, aligns well with DL advancements like the Unet++ framework, underscoring its significance in this field. In applying DL to remote sensing images, a critical challenge is the accurate extraction and discerning prioritization of salient features, a task that mirrors human cognition. Among numerous visual stimuli, the human brain adeptly emphasizes essential details. This parallels the role of attention mechanisms in neural networks. Such mechanisms allow models to allocate varying significance to different data facets, enhancing both feature extraction accuracy and model interpretability. As a result, this has led to the development of attention-augmented methods, which can be largely categorized into spatial-wise, channel-wise, and mixed attention approaches.

1) Spatial-Wise Attention: Spatial attention mechanisms were introduced to guide neural networks in pinpointing specific regions within an image that warranted emphasis. This was particularly crucial in remote sensing, where understanding spatial distributions and patterns was paramount. The mechanism worked by transforming the original data into a latent space, preserving only the task-relevant patterns. A notable method in this context was PGA-SiamNet [16]. It focused on emphasizing regions in paired images that differed and downplaying identical areas. Building on such concepts, difference-enhancement dense-attention modules [17] were integrated, especially following the U-Net backbone. These modules produced attention masks at different levels, further enhancing the model's resilience to noise. Methods such as [6] and [29] leveraged the Transformer's capability to capture long-range information, modeling the global spatial context for improved results. However, although spatial attention provided a robust understanding of image space, it did not always offer a granular view of the diverse channels of information inherent in remote sensing imagery. This realization prompted the exploration of channel attention.

2) Channel-Wise Attention: This mechanism focuses on enabling neural networks to determine which feature channel requires more emphasis. The IFN [15] was conceptualized to enhance internal compactness by introducing channel attention across every encoder layer. Later developments, like the UNet++ backbone [14], shifted towards deriving a multitude of feature maps from decoders. In this context, the Ensemble Channel Attention Module was devised to refine the most salient features across various levels. As spatial attention excels in capturing local dependencies and channel attention adeptly identifies feature-wise importance, both mechanisms have their constraints. Integrating these approaches not only capitalizes on their strengths, but also addresses inherent weaknesses. This rationale sets the stage for our exploration of mixed attention.

3) Mixed Attention: Building upon spatial and channel attention mechanisms, the mixed attention approach integrates these into a unified framework. The Convolutional Block Attention Module [30], for instance, enhances feature maps for complex relationship analysis in change detection [19]. Similarly, DCA-Det [18] employs mixed attention modules for object-level change refinement. These developments, coupled with domain adaptation techniques like HighDAN [31], underscore the dynamic progress in DL for remote sensing change detection. To adeptly capture global patterns from an image without incurring significant computational demands, STANet [22] divided images into patches and leveraged a spatial-temporal attention module, treating temporal information as channels. It refined features by computing attention weights between pixel pairs. Likewise, BIT [20] employed a semantic tokenizer to convert image pairs into patches, enhancing feature representation across both spatial and channel dimensions.

C. Transformer-Based Deep Learning Methods

The Transformer architecture, first presented in [32], has since become the benchmark in natural language processing. Building on this success, Dosovitskiy et al. [33] extended the standard Transformer to computer vision, achieving notable results in image recognition. Likewise, in semantic segmentation, Transformers have made significant advancements. Strudel et al. [34] presented the first pure Transformer-based method, yet its high computational cost remained a challenge. Xie et al. [35] addressed this by refining the self-attention mechanism and resizing the feature map, striking a balance between efficiency and performance. Although the self-attention mechanism is often credited as the pivotal component in Transformers, the very recent studies [36] indicate that replacing the self-attention module with alternative aggregation operations still yields competent models in computer vision tasks. For change detection, MATU [6] combined CNNs and Transformers to learn features efficiently. Meanwhile, ChangeFormer [29] employed a pure Transformer within the Siamese structure specifically for this purpose. Building on these advancements, SpectralGPT [37] was introduced for processing spectral remote sensing (RS) images using a 3-D generative pretrained transformer. This model is designed to analyze diverse and complex RS data, improving performance in tasks like change detection. However, most of these methods prioritize long-range dependencies in individual inputs, often overlooking correlations between image pairs. Our research aims to bridge this gap by focusing on both interimage dynamics and intraimage correlations.

III. Methodology

This section provides an in-depth explanation of our novel SBA-PN. As illustrated in Fig. 2, the SBA-PN framework introduces an innovative approach to bitemporal change detection. The architecture effectively combines two essential components: 1) a multistage Siamese encoder, of which each stage consists of several Spatial Optimal Pooling (SOP) blocks aiming to extract multiscale feature pairs from bitemporal images and 2) a multistage biattention decoder, which consists of biattention

Algorithm 1: Inference of SPA-PN Model.

```
Input: $\mathbf{I} = \{(\mathbf{I}_1, \mathbf{I}_2)\}$ (a pair of registered images)
Output: $\hat{\mathbf{C}}$ (a binary change map)
#Step 1: extract feature pairs by Encoder
for $i$ in $\{1, 2\}$ do
    $\{\mathbf{F}_i^s\}_{s \in S} = \text{Encoder}(\mathbf{I}_i)$
end for
#Step 2: use Bi-Attention Decoder to refine the distance maps
$\mathbf{D}^S = |\mathbf{F}_1^S - \mathbf{F}_2^S|$ (initial distance map at stage S)
for $s = S - 1$ down to 1 do
    $\mathbf{M}^s = \text{Bi-Attention}(\mathbf{F}_1^s, \mathbf{F}_2^s, \mathbf{D}^s)$
    $\mathbf{D}^{s-1} = \text{channel\_deviation\_pooling}(\mathbf{M}^s)$
end for
#Step 4: obtain change mask by the prediction head
$\hat{\mathbf{C}} = \text{Detection\_Head}(\mathbf{D}^1)$
```

modules and CDP blocks to progressively generate change maps from deep to low. Given a pair of bitemporal images at the preevent time and the postevent time, denoted by $\mathbf{I}_1 \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{I}_2 \in \mathbb{R}^{H\times W\times C}$, respectively, where W indicates the width, H indicates the height, and C is the number of channels. The primary objective of a change detection algorithm is to estimate a binary change map $\mathbf{C} \in \mathbb{R}^{H\times W}$, which can be defined as follows:

$$\mathbf{C} = \{c_{wh} \in \{0, 1\}, 1 \leq w \leq W, 1 \leq h \leq H\}. \tag{1}$$

Particularly, we denote the estimated change map as $\hat{\mathbf{C}} = \{\hat{c}_{wh}\} \in \mathbb{R}^{W\times H}$. In the following subsections, we detail each component of the proposed methodology and highlight the innovations that distinguish our approach.

A. Encoder With SOP Blocks

The encoder is constructed in a multistage manner with S stages, each of which contains an embedding block and an SOP block.

1) Embedding Block: At the beginning of an encoding stage, the input is first processed by an embedding block to reduce its dimension as a downsampling operation, which is implemented by a convolution layer. We have

$$\mathbf{X}_i = \text{Embedding}\left(\mathbf{I}_i\right). \tag{2}$$

2) SOP Block: The embedded features $\mathbf{X}_i$, for $i \in \{1, 2\}$, are processed through L consecutive SOPs. Each SOP comprises an aggregation layer, at the core of which is an aggregator that contextualizes and refines spatial visual features. This operation is captured by

$$\mathbf{Y}_i = \text{Aggregator}\left(\text{Norm}\left(\mathbf{X}_i\right)\right) + \mathbf{X}_i \tag{3}$$

where $\text{Norm}(\cdot)$ denotes a normalization function, based on the Group Normalization method [38]. The output of the aggregation layer is denoted as $\mathbf{Y_i}$ for $i \in \{1, 2\}$. Various aggregators have been investigated in the literature [36]. A self-attention based aggregator has been widely used for computer vision tasks by treating $\mathbf{X}_i$ as a set of subimage patches, and in this case (3) can be viewed as a Transformer block. Besides, spatial multilayer perceptron (MLP), convolution and pooling-based aggregators have also been explored. It is noteworthy that, although the pooling-based aggregator is straightforward in its approach, it demonstrates a comparable modeling capability with significantly fewer parameters, thereby reducing the model complexity [36].

Motivated by this, we aim to devise an optimal strategy by taking the advantages of this pooling based scheme. In this study, pooling operations summarise the pixel-wise patterns in a given region or an observation window. Formally, we denote by $\mathbf{p}_n \in \mathbb{R}^D$, where $n = 1, \ldots, N$, the vectors characterising the n^{th} location (i.e., pixel) in a region. Here, N represents the total number of pixels in the given region. A general average pooling and maximum pooling can then be adopted accordingly.

Average pooling computes the mean of $\mathbf{p}_n$

$$y^d = \frac{1}{N}\sum_{i=1}^{N} \mathbf{p}_i^d, \forall d \tag{4}$$

where p_i^d is the value of the dth dimension in p_n, and $\mathbf{y} = \{y^d\}_{d=1:D} \in \mathbb{R}^D$ is the pooled result from all locations.

Maximum pooling operates by identifying the maximum value with a given observation window

$$y^d = \max_i \left(p_i^d\right), \forall d. \tag{5}$$

Note that the average pooling treats all elements equally which may introduce redundant patterns for change detection. Conversely, the maximum pooling focuses only on the dominant element, which could miss critical change detection hints. Hence, to achieve an optimal and balanced pooling scheme for change detection, we devise a learnable pooling strategy–SOP.

SOP weights the top K elements in a data-driven manner for a given observation window

$$y^d = \sum_{k=1}^{K} w_k \cdot \text{k-}\max_i \left(\mathbf{p}_i^d\right), \forall d \tag{6}$$

where $w_k > 0$ and $\sum_{k=1}^{K} w_k = 1$. $\text{k-}\max_i$ is an operator that identifies the k greatest values. Note that the coefficients w_k are learnable weights. The SOP can be used to imitate convolution and average pooling with $K = N$, or approximate maximum pooling with $K = 1$. As a result, (3) can be written as

$$\mathbf{Y}_i = \text{SOP}\left(\text{Norm}\left(\mathbf{X}_i\right)\right) + \mathbf{X}_i. \tag{7}$$

Next, the linear transformation layer, which is also known as a feed-forward layer, is adopted following a common practice in visual Transformers [33]. In detail, it consists of two MLPs

$$\mathbf{Z}_i = \text{FeedForward}\left(\mathbf{Y}_i\right) = \sigma\left(\text{Norm}\left(\mathbf{Y}_i\right)\mathbf{W}_1\right)\mathbf{W}_2 + \mathbf{Y}_i \tag{8}$$

where σ is a nonlinear activation function. $\mathbf{W}_1$ and $\mathbf{W}_2$ are matrices containing learnable weights.

By adopting a number of SOPs, the final output is utilized as the input of the next encoding stage to construct deeper feature representations. The obtained encoded feature pairs from each stage is denoted by $\mathbf{F}_1^s$ and $\mathbf{F}_2^s$, where s is the stage index.

B. Biattention Decoder

The purpose of the decoder in change detection is to understand the paired encoding features and formulate their difference and the corresponding change map.

Similar to the encoder, there are multiple stages $s = S, \ldots, 1$ in the decoding process. Each decoding stage contains G biattentions and a CDP module. In the sth decoding stage, its input is the feature pair $\mathbf{F}_1^s$ and $\mathbf{F}_2^s$ from the corresponding sth encoding stage and a distance map $\mathbf{D}^s$. It then formulates a distance map $\mathbf{D}^{s-1}$ for the next decoding stage. Particularly, the distance map $\mathbf{D}^S$ to the initial decoding stage S is obtained as

$$\mathbf{D}^S = \left|\mathbf{F}_1^S - \mathbf{F}_2^S\right| \tag{9}$$

where $|\cdot|$ indicates an element-wise operator for absolute values.

1) Biattention Module: Drawing inspiration from the self-attention mechanism introduced in [32], we present a refined biattention layer that marries the multihead biattention (MBA) design with linear projections. For each layer l where $l = 1, \ldots, G$, and for every head i with $i = 1, \ldots, \Gamma$, a query-key-value triplet, namely $\mathbf{Q}_{i,l}^s$, $\mathbf{K}_{i,l}^s$, and $\mathbf{V}_{i,l}^s$, is formulated as

$$\mathbf{Q}_{i,l}^s = \mathbf{F}_1^s \mathbf{W}^{\mathbf{Q}^s}{}_{i,l};\ \mathbf{K}_{i,l}^s = \mathbf{F}_2^s \mathbf{W}^{\mathbf{K}^s}{}_{i,l};\ \mathbf{V}_{i,l}^s = \mathbf{O}_{l-1}^s \mathbf{W}^{\mathbf{V}^s}{}_{i,l} \tag{10}$$

where $\mathbf{W}^{\mathbf{Q}^s}{}_{i,l}$, $\mathbf{W}^{\mathbf{K}^s}{}_{i,l}$, $\mathbf{W}^{\mathbf{V}^s}{}_{i,l} \in \mathbb{R}^{c_s \times d}$ are matrices containing learnable parameters. Note that $\mathbf{F}_1^s \in \mathbb{R}^{H_s W_s \times C_s}$ and $\mathbf{F}_2^s \in \mathbb{R}^{H_s W_s \times C_s}$ are reformulated as $H_s W_s$ vectors, $\mathbf{O}_{l-1}^s$ is the output of the previous biattention layer. Particularly, the input to the initial biattention layer is the distance map: $\mathbf{O}_0^s = \mathbf{D}^s$.

In addition, to consider the spatial relations of these vectors, position information is introduced to the input. Absolute position encoding is added to the triplet before being fed into the first biattention layer in each decoding stage

$$\mathbf{F}_1^s = \mathbf{F}_1^s + \mathbf{P_Q}^s;\ \mathbf{F}_2^s = \mathbf{F}_2^s + \mathbf{P_K}^s;\ \mathbf{O}_0^s = \mathbf{O}_0^s + \mathbf{P_V}^s \tag{11}$$

where $\mathbf{P_Q}^s$, $\mathbf{P_K}^s$, $\mathbf{P_V}^s \in \mathbb{R}^{H_s W_s \times C_s}$ are positional embeddings. The ith attention head models the global relations between the triplet as

$$\text{Head}_{i,l}^s\left(\mathbf{Q}_{i,l}^s, \mathbf{K}_{i,l}^s, \mathbf{V}_{i,l}^s\right) = \sigma\left(\frac{\mathbf{Q}_{i,l}^s {\mathbf{K}_{i,l}^s}^\top}{\sqrt{d}}\right)\mathbf{V}_{i,l}^s \tag{12}$$

where σ denotes a softmax activation function on channel dimension and d is the channel dimension of the triplet.

We compute Γ attention heads simultaneously and concatenate them to form a complete MBA

$$\begin{aligned}\mathbf{A}_l^s &= \text{MBA}\left(\text{Head}_{1,l}^s, \ldots, \text{Head}_{\Gamma,l}^s\right)\\ &= \left(\text{Head}_{1,l}^s \oplus \ldots \oplus \text{Head}_{\Gamma,l}^s\right)\mathbf{W}^{\mathbf{A}_l^s}\end{aligned} \tag{13}$$

where $\mathbf{W}_{\mathbf{A}_l^s} \in \mathbb{R}^{\Gamma d \times c_s}$ is a matrix for linear projection and $\oplus$ is a concatenation operator along the channel dimension. Finally, the decoding result goes through an MLP of two layers

$$\mathbf{O}_l^s = \text{MLP}\left(\mathbf{A}_l^s\right). \tag{14}$$

By adopting G consecutive biattention layers, the output is ready to be fed into the next stage of the decoder. A short-cut connection with the next corresponding encoding stage's output is introduced as follows:

$$\mathbf{M}^s = \text{upsample}\left(\mathbf{O}_G^s\right) \oplus \left|\mathbf{F}_1^{s-1} - \mathbf{F}_2^{s-1}\right| \tag{15}$$

where the upsampling operator is necessary to upsample $\mathbf{M}^s$ to a spatial dimension of $H_{s-1} \times W_{s-1}$, which matches the dimension of $\mathbf{F}_1^{s-1}$ and $\mathbf{F}_2^{s-1}$.

2) CDP Module: The output channel dimension needs to be aligned with the next stage of the decoder. Unlike traditional methods that use spatial convolutions, we found that incorporating additional channel attention can further emphasize patterns relevant to change detection. This is achieved by selectively emphasizing informative channels and filtering out redundancy. Motivated by this observation, we introduce a CDP mechanism on the decoded feature map $\mathbf{M}^s$ to obtain the results of the sth decoding stage

$$\begin{aligned}\mathbf{D}^{s-1} &= \text{CDP}\left(\mathbf{M}^s\right)^\top,\\ \text{CDP}\left(\mathbf{M}^s\right) &= \text{CDP-Pool}\left(\mathbf{H}_{\text{CDP}}^s\right)\mathbf{V}_{\text{CDP}}^{s\top},\\ \mathbf{H}_{\text{CDP}}^s &= \sigma\left(\frac{\mathbf{Q}_{\text{CDP}}^{s\top}\mathbf{K}_{\text{CDP}}^s}{\sqrt{d}}\right)\end{aligned} \tag{16}$$

where $\mathbf{Q}_{\text{CDP}}^s$, $\mathbf{K}_{\text{CDP}}^s$, and $\mathbf{V}_{\text{CDP}}^s \in \mathbb{R}^{H_{s-1}W_{s-1} \times E_s}$ are matrices obtained from $\mathbf{M}^s$ with learnable linear projections; and $\mathbf{H}_{\text{CDP}}^s \in \mathbb{R}^{E_s \times E_s}$, which is the product of transposed $\mathbf{Q}_{\text{CDP}}$ and $\mathbf{K}_{\text{CDP}}$, can be viewed as a feature heatmap. CDP-Pool$(\cdot)$ denotes the function to select C_{s-1} rows from $\mathbf{H}_{\text{CDP}}^s$ as a new matrix, which contains the most informative patterns. Formally, we rank the rows in $\mathbf{H}_{\text{CDP}}^s$ by computing row-wise standard deviation, and a row with high standard deviation indicates its informative nature.

C. Change Detection Head

The decoder progressively restores the output feature map for change detection back to the original image size. A prediction head is further adopted to predict changed results based on the final feature map $\mathbf{D}^1$. Specifically, the prediction head generates a probability map $\hat{\mathbf{C}} \in \mathbb{R}^{H \times W}$ to estimate $\mathbf{C}$

$$\hat{\mathbf{C}} = \sigma(g\left(\text{upsample}\left(\mathbf{D}^1\right)\right)) \tag{17}$$

where σ is a pixel-wise softmax function along the channel dimension and g is a convolutional layer.

D. Contrastive Pixel-Wise Supervision

The success of DL techniques is closely tied to the growth in the depth of neural networks. However, traditional training strategies typically supervise the neural network at its final layer and use back-propagation to adjust the remaining layers. It presents a significant challenge in optimizing the intermediate layers. Hence, recent research [19], [39], [40] has introduced the concept of applying auxiliary supervision directly to the shallow layers. However, this strategy has its limitations in the high-level semantic supervision, which is inherently task-focused, often conflicts with the common observation that shallow layers primarily learn low-level features.

A potential solution could be put forth in a contrastive deep supervision framework [41], which supervises the intermediate layers using augmentation-based classifiers. This method, however, was primarily designed for image-level classification tasks, and its effectiveness could diminish when applied to our change detection one, which is a dense prediction task. Therefore, we devise a self-supervised learning scheme for dense predictions, which operates in a pixel-wise manner to overcome the limitations of existing methods as a more optimal solution.

The proposed optimization scheme focuses on the outputs from each encoding stage, and we denote $\mathbf{f}$ as the vector of a pixel in the changed area of $\mathbf{F}_1^s$. Its associated positive vector $\mathbf{f}^+$ is a sampled pixel point from the neighboring location on $\mathbf{F}_1^s$; whilst its negative associated vectors $\mathbf{f}^-$ are a set of pixels in the changed area on $\mathbf{F}_2^s$. As revealed by recent studies [42], [43], a large set of negatives is critical in contrastive representation learning. To this end, the pixel-wise contrastive loss is defined as

$$\mathcal{L}_{\text{contra}} = \sum_{\mathbf{f}} -\log \frac{\exp\left(\mathbf{f}\cdot\mathbf{f}^+/\tau\right)}{\exp\left(\mathbf{f}\cdot\mathbf{f}^+\right) + \sum_{\mathbf{f}^-}\exp\left(\mathbf{f}\cdot\mathbf{f}^-/\tau\right)} \tag{18}$$

where τ denotes a temperature hyperparameter.

E. Training Loss

As change detection can be viewed as a binary pixel-wise classification, the binary cross entropy (BCE) loss is adopted to optimize our network

$$\mathcal{L}_{\text{bce}} = \frac{1}{WH}\sum_{w,h}\ell(\hat{c}_{wh}, c_{wh}),$$
$$\ell(\hat{c}_{wh}, c_{wh}) = -c_{wh}\log(\hat{c}_{wh}) + (1-c_{wh})\log(1-\hat{c}_{wh}). \tag{19}$$

To this end, the overall loss function is a linear combination of the BCE loss and the pixel-wise contrastive loss

$$\mathcal{L}_{\text{total}} = 0.5 \times \mathcal{L}_{\text{bce}} + 0.5 \times \mathcal{L}_{\text{contra}}. \tag{20}$$

IV. Experiments

A. Datasets

To demonstrate the effectiveness of our method, comprehensive experiments are conducted on three commonly used benchmark datasets, LEVIR-CD [22], SYSU-CD [19], and HRCUS-CD [46].

1) LEVIR-CD: The LEarning, VIsion, and Remote sensing Change Detection (LEVIR-CD) dataset distinguishes itself on building change detection. Encompassing 637 pairs of high-resolution remote sensing images, each image exhibits a dimension of 1024 × 1024 at spatial resolution of 0.5 m. In the public version of the dataset, these images are systematically divided into distinct, nonoverlapping sub-images of 256 × 256. As stated in [22], the dataset is partitioned into training, validation, and testing subsets, comprising 7120, 1024, and 2048 image pairs, respectively.

2) SYSU-CD: The Sun Yat-Sen University Change Detection (SYSU-CD) dataset emerges as a notable public resource for bitemporal image CD, illuminating the dynamic landscape of Hong Kong between 2007 and 2014. It comprises 20 000 pairs of aerial images, each meticulously captured at a 0.5-m granularity and displayed at a resolution of 256 × 256. The captured changes span a diverse range: From the rise of new urban infrastructures and roadway expansions to fluctuations in vegetation and coastal developments. Structured with precision, the dataset adopts a 6:2:2 distribution ratio, designating 12 000 pairs for training, 4000 for validation, and 4000 for testing, thus ensuring a robust and balanced experimental setup.

3) HRCUS-CD: The High-Resolution Complex Urban Scene Change Detection (HRCUS-CD) dataset includes 11 388 pairs of 256 × 256 pixel high-resolution remote sensing images with a 0.5-m spatial resolution, containing over 12 000 annotated change instances. Originating from Zhuhai, China, the dataset captures two primary areas: 1) the Urban Built-up Area (2019–2022) with minimal changes and 2) the Rural and Developing Urban Area (2010–2018) with notable urban development.

B. Implementation Details

1) Training Setup: We implemented our models using PyTorch and trained them on a single NVIDIA Tesla V100 GPU. We applied standard data augmentation techniques to the input images, which include flipping, rescaling, cropping, and Gaussian blurring. The batch size is set to 8. The model was optimized using stochastic gradient descent with momentum. We set the momentum to 0.99 and the weight decay factor to 0.0005. The learning rate is initially set at 0.01 and linearly decays to 0 over the course of 200 training epochs. In (18), we set the temperature τ to 0.07, following the setting in [43]. Validation is performed after each training epoch, and the best-performing model on the validation set is subsequently used for the evaluation on the test set.

2) Size Variations: This section outlines the various configurations of our SBA-PN model, highlighting its modularity and scalability across diverse settings, as follows.

1) **SBA-PN/S** (Small Size Configuration): The encoder employs the MetaFormer architecture [36] as its backbone. MetaFormer features four stages, with each stage reducing the spatial size of its input to one-fourth of its original dimensions. For this configuration, we use the "S24" version, which means the model includes 24 MetaFormer blocks. The output channel dimensions from this configuration are [64, 128, 320, 512]. The feed-forward layer integrates an MLP with an expansion ratio of 4. As mentioned in Section III, our adaptation of the MetaFormer involves replacing its token mixer with our SOPs, where the pooling operation in our SOP has a kernel size of 3 × 3, and the optimal value k is 5. The decoder incorporates six biattention layers for each stage. Absolute position encoding is used in our biattention mechanism, where these position encodings are embedded with each input. These are learnable parameters with values between 0 and 1 and are as long as the input sequence length. The multihead biattention mechanism features eight heads, with each head having a channel dimension d of 8. The feed-forward layers mirror the MetaFormer's specifications with an expansion ratio of 4. The prediction head consists of two convolutional

TABLE I
COMPARISON BETWEEN OUR PROPOSED SBA-PN AND EXISTING METHODS IN TERMS OF CHANGE DETECTION PERFORMANCE (%) AND MODEL COMPLEXITY (I.E., PARAMS. AND FLOPS) ON LEVIR-CD, SYSU-CD, AND HRCUS-CD

Model	**Params.**	**Flops**	**LEVIR-CD**				**SYSU-CD**				**HRCUS-CD**			
	(M)	**(G)**	**Pre.**	**Rec.**	**F1**	**IoU**	**Pre.**	**Rec.**	**F1**	**IoU**	**Pre.**	**Rec.**	**F1**	**IoU**
FC-EF [5]	1.29	2.92	86.91	80.17	83.40	71.53	74.32	75.84	75.07	60.09	64.99	61.15	63.01	45.99
FC-DIFF [5]	1.93	4.55	87.53	83.31	85.37	74.47	**89.13**	61.21	72.57	56.96	64.29	67.76	65.98	49.23
FC-CONC [5]	1.75	3.99	91.99	76.77	83.69	71.96	82.54	71.03	76.35	61.75	53.95	66.95	59.75	42.61
STANet [22]	16.93	6.58	80.99	91.21	85.79	75.12	82.36	74.30	78.12	64.10	67.29	67.36	67.32	50.74
Bit-CD [20]	21.67	10.06	91.95	88.57	90.23	82.19	79.04	76.71	77.86	63.75	73.29	67.18	70.11	53.97
ChangeFormer [29]	267.90	129.27	91.53	88.86	90.17	82.10	84.99	70.93	77.33	63.04	67.68	64.45	66.55	49.86
SwinSUNet [44]	40.95	11.19	90.51	89.72	90.11	82.00	83.05	72.67	77.51	63.28	66.78	64.62	65.68	48.90
AMTNet [45]	24.67	21.56	91.82	89.71	90.76	83.08	78.20	77.92	78.06	64.01	66.14	65.91	66.03	49.28
AERNET [46]	25.36	12.82	89.97	91.59	90.78	83.11	78.33	78.26	78.30	64.33	77.17	**77.05**	**77.11**	**62.75**
SBA-PN/S	25.75	13.69	91.23	90.46	90.84	83.22	82.13	77.86	79.94	66.58	74.87	74.50	74.69	59.60
SBA-PN/L	43.54	18.04	**93.10**	**91.63**	**92.36**	**85.80**	83.43	**81.26**	**82.33**	**69.97**	**78.55**	75.17	76.82	62.37

[1] The best performance is highlighted in bold.

layers with Group Normalization (GN). These layers have output channels of 32 and 2, respectively, and use a kernel size of 3.

2) **SBA-PN/L** (Large Size Configuration): For this variant, the encoder uses the "M36" configuration of the MetaFormer architecture, indicating a larger model with 36 MetaFormer blocks. The decoder in this configuration includes eight biattention layers for each stage. The multihead biattention mechanism here comprises 16 heads, allowing for more complex attention patterns.

3) Evaluation Metrics: Change detection can be understood as a pixel-wise binary classification challenge. For evaluating the prediction outcomes, we primarily rely on metrics such as precision, recall, F1 score, and Intersection over Union (IoU). Precision provides insights into the accuracy of positive predictions, and recall denotes the proportion of true positives that are correctly identified. The F1 score harmoniously combines precision and recall, serving as a balanced metric. In addition, IoU offers a measure of the overlap between the predicted change map and the ground truth. In this study, the F1 score is our predominant evaluation metric due to its resilience against the skewed distributions often observed between changed and unchanged classes. The metrics are concisely defined by the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{21}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{22}$$

$$\text{F1 Score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \tag{23}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{24}$$

where TP, FP, FN, and TN represent true positives, false positives, false negatives, and true negatives, respectively.

C. Comparison With SOTA Methods

To highlight the performance of our proposed method, we compared it with in total six state-of-the-art (SOTA) methods known for their robust performance: FC variations [5], STANet [22], BIT [20], ChangeFormer [29], SwinSUNet [44], AMTNet [45], and AERNET [46]. In the following subsection, we provide a detailed discussion, presenting both quantitative results and qualitative examples.

1) Quantitative Analysis: Our proposed methods, SBA-PN/L and SBA-PN/S, demonstrate superior performance across various datasets when compared to existing methods, as shown in Table I. The best performance is highlighted in bold and the second-best performance is underlined for clarity.

Starting with the LEVIR-CD dataset, our methods achieve remarkable F1-scores of 92.05 for SBA-PN/L and 91.03 for SBA-PN/S, along with IoU scores of 84.77 and 83.27, respectively. Notably, SBA-PN/L surpasses the best existing method, Bit-CD, by 1.82% in F1-score and 2.58% in IoU. Similarly, on the SYSU-CD dataset, SBA-PN/L and SBA-PN/S record F1-scores of 82.78 and 80.31, respectively, with corresponding IoU scores of 71.06 and 67.26. Here, SBA-PN/L outperforms the top existing method, STANet, by 4.66% in F1-score and 6.96% in IoU.

In addition, our analysis of the HRCUS-CD dataset reveals that SBA-PN/L achieves the highest precision score of 78.55%, demonstrating its strong capability in accurately detecting changes in complex urban settings with high-resolution imagery. While SBA-PN/L does not achieve the top IoU score, its second-place ranking at 62.37% still showcases a commendable performance in identifying true changes, underlining the importance of a balanced approach in change detection for varied environments.

In terms of computational efficiency, our SBA-PN models are also notable for their reduced complexity. The SBA-PN/S model requires only 13.69 GFLOPs, and the SBA-PN/L model operates at 18.04 GFLOPs. This is significantly lower compared to other high-performing models such as ChangeFormer, which demands 129.27 GFLOPs. This stark contrast in FLOPs demonstrates the efficiency of our models, making them suitable for scenarios where computational resources are limited. Despite their lower FLOPs, both SBA-PN models maintain superior performance metrics, underscoring their ability to provide an optimal balance between accuracy and computational load.

However, it is important to note that in terms of precision, our SBA-PN method exhibits lower scores compared to some existing methods on the SYSU-CD dataset. These methods, while achieving higher precision, tend to have significantly reduced

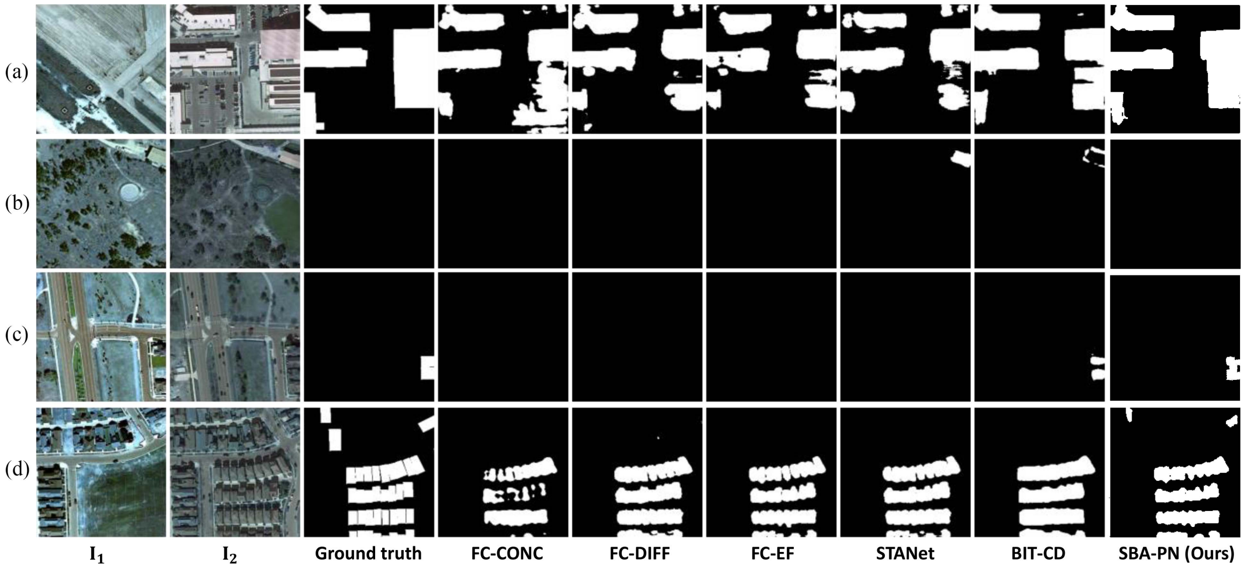

Fig. 3. Qualitative results of various methods on four image pairs (i.e., (a)–(d)) from the LEVIR-CD test set. The first two columns are the input image pair, the third column is the ground truth change map, the fourth to the ninth columns show the results of FC-CONC [5],..., BIT-CD [20] and our proposed method SBA-PN.

recall scores, leading to an overall lower F1-score than SBA-PN. This suggests a potential bias towards unchanged pixels in these methods, likely due to the inherent imbalance between changed and unchanged areas in the datasets.

Finally, when comparing the three datasets, LEVIR-CD appears to be the least challenging. This is indicated by the higher performance metrics achieved by our methods on this dataset. The less challenging nature of LEVIR-CD can be attributed to its sole focus on building changes. In contrast, SYSU-CD and HRCUS-CD encompasses a more diverse range of changes and a larger number of image pairs, thereby presenting a more complex scenario for change detection.

2) Qualitative Comparison: Fig. 3 provides a visual comparison of different methods applied to 4 sample pairs in the LEVIR-CD dataset, where true positive pixels (changed areas) are shown in white and true negative pixels (unchanged areas) are shown in black. From our observations, the proposed SBA-PN outperforms other methods in a number of aspects.

First, SBA-PN with refined discriminative features considering the global context effectively reduces false negatives, which often arise when distinct objects appear under specific lightning or weather conditions (e.g., roofs under shadows). Conversely, in Fig. 3(a), most existing methods incorrectly classify the gray rectangular stripes on the building's roof as unchanged pixels, leading to an inconsistent result for the rightmost rectangular building. Likewise, in Fig. 3(c), existing methods fail to detect changes at the right corner due to similarities in texture. Second, SBA-PN effectively mitigates irrelevant changes induced by seasonal variations or modifications in the appearance of land cover elements.

Fig. 4 presents a qualitative comparison of different methods applied to the SYSU-CD dataset. Overall, both BIT-CD and SBA-PN demonstrate superior capability to detect entire changes compared to other methods, which often exhibit patchy or noisy remnants. However, our SBA-PN method has a lower false positive rate than BIT-CD. For example, in Fig. 3(d), a large area of land surrounding the orange object is incorrectly identified as a changed region by BIT-CD. In contrast, our method is not only comprehensive for large-scale structural patterns, but also precise for fine-grained patterns.

D. Ablation Study

This section provides a detailed ablation study aimed at understanding the influence of individual components within our proposed method, as shown in Table II.

SOP: In comparison to the default aggregation operation in PoolFormer that employs an average pooling strategy, our SOP demonstrated enhancements in F1-scores, registering increases of 0.89 and 1.67 on the two datasets. This improvement was achieved with only a marginal increase in computational cost. These findings underscore the pivotal role our customized backbone and SOP play in enhancing change detection performance. Visually, it is apparent that SOP helps to globally detect changes. From Fig. 5(a), it is evident that the absence of SOP results in false positive patches at the bottom of the change map. While these may appear as changes locally, they are not considered changes on a global scale.

Biattention: To assess the effectiveness of our biattention mechanism, we compared it with the widely-used self-attention mechanism. The findings indicated that our biattention mechanism outperformed the standard self-attention approach, recording F1-score improvements of 0.52 and 0.63 on the two datasets, respectively. As illustrated in Fig. 5(a) and (b), the biattention module appears to smooth out the change detection results by removing sparkling noise. This highlights the critical role of the biattention mechanism in our model, enabling a more focused exchange of information between paired feature sets.

CDP: Omitting the contrastive component from the SBA-PN model resulted in slight reductions in both the number of parameter and the number of FLOPs, indicating a simpler model

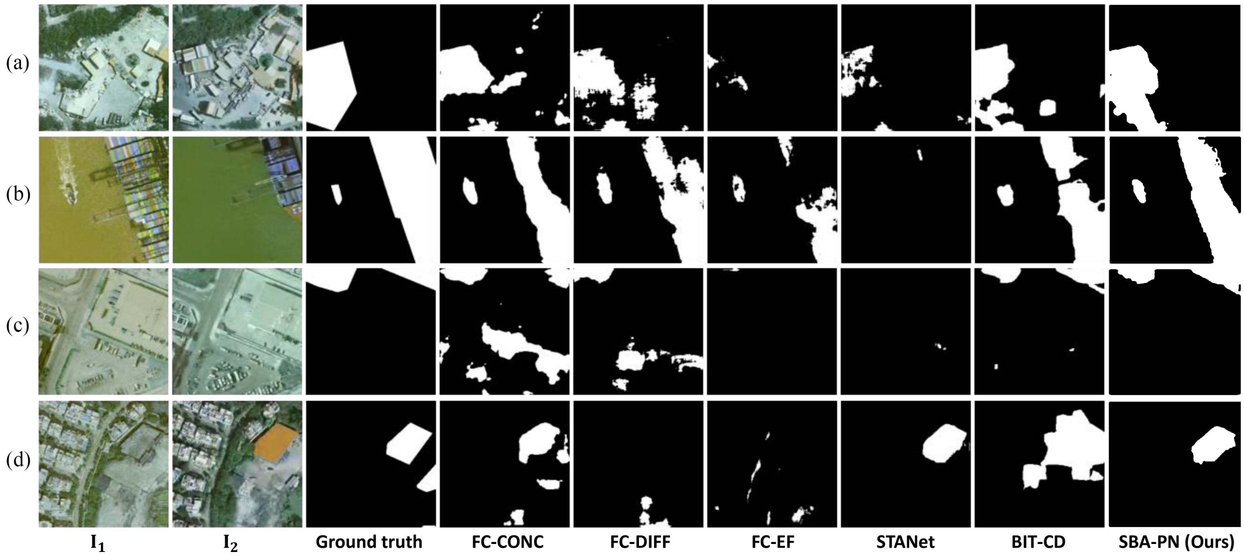

Fig. 4. Qualitative results of various methods on four image pairs (i.e., (a)–(d)) from the SYSU-CD test set. The first two columns are the input image pair and the third column is the ground truth change map. The fourth to the ninth columns display the prediction outcomes of FC-CONC [5],..., BIT-CD [20] and our proposed method SBA-PN, respectively.

TABLE II
EXPERIMENTAL RESULTS OF ABLATION ON LEVIR-CD AND SYSU-CD

Model	SOP	Bi-attention	CDP	Params. (M)	Flops (G)	LEVIR-CD				SYSU-CD			
						Pre.	Rec.	F1	IoU	Pre.	Rec.	F1	IoU
SBA-PN	✓	✓	✓	25.75	13.69	**91.23**	**90.46**	**91.03**	**83.27**	82.13	**77.86**	**80.31**	**67.26**
SBA-PN (w/o SOP)	✗	✓	✓	24.88	10.12	90.37	89.86	90.14	81.98	81.96	75.98	78.64	65.30
SBA-PN (w/o Bi-Attention)	✓	✗	✓	13.41	6.67	87.78	86.32	86.33	77.86	81.20	72.68	76.54	63.10
SBA-PN (w/o CDP)	✓	✓	✗	26.31	12.22	90.02	89.98	90.79	82.88	**83.01**	75.56	79.42	66.31
SBA-PN (w/o Contrastive)	✓	✓	✓	25.75	13.14	90.74	89.75	90.59	82.73	82.29	75.19	79.32	66.81
SBA-PN (w/o SOP + Bi-Attention)	✗	✗	✓	18.03	10.14	87.01	84.28	85.42	76.54	80.13	71.69	75.77	62.24
SBA-PN (w/o SOP + CDP)	✗	✓	✗	28.03	9.96	89.33	89.15	88.37	80.67	81.94	73.56	74.24	64.13
SBA-PN (w/o Bi-Attention + CDP)	✓	✗	✗	12.33	17.11	87.02	86.03	85.79	77.29	81.01	72.01	75.98	62.65

The best performance is highlighted in bold.

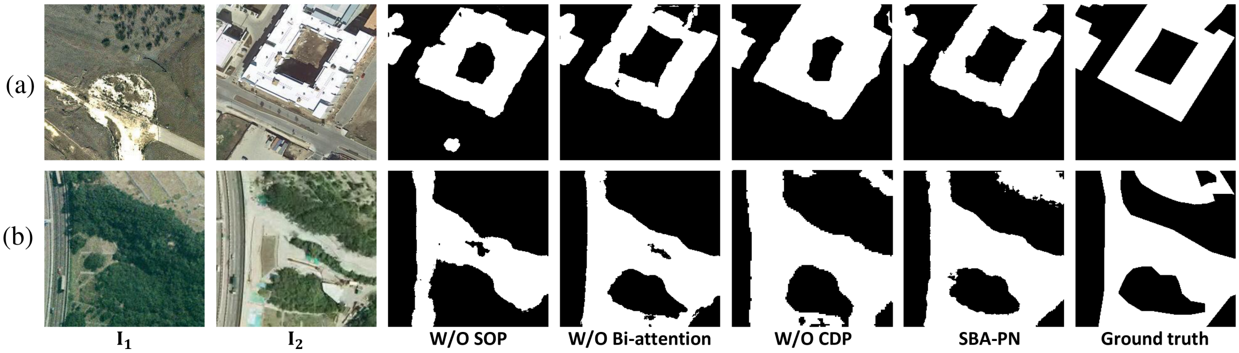

Fig. 5. Visualized comparisons of ablation studies: (a) selected samples from the LEVIR-CD dataset and (b) selected samples from the SYSU-CD dataset.

with potentially lower computational requirements. However, this change negatively impacted performance across all metrics for the LEVIR-CD and SYSU-CD datasets. This performance drop signifies the importance of the contrastive component on the SBA-PN model's efficiency and effectiveness, especially on the SYSU-CD dataset.

Pixel-wise contrastive: When comparing the CDP setting with traditional pointwise convolution, CDP proved superior in terms of both efficiency and effectiveness. This advantage arises from the CDP's approach, which identifies key features and ignoring less relevant semantic information across the channel dimension.

Moreover, we validated the effectiveness of combining various modules by testing three combinations: "SOP + Bi-attention," "SOP + CDP," and "Bi-attention + CDP." As indicated in Table II, the results show that the 'SOP + Bi-attention'

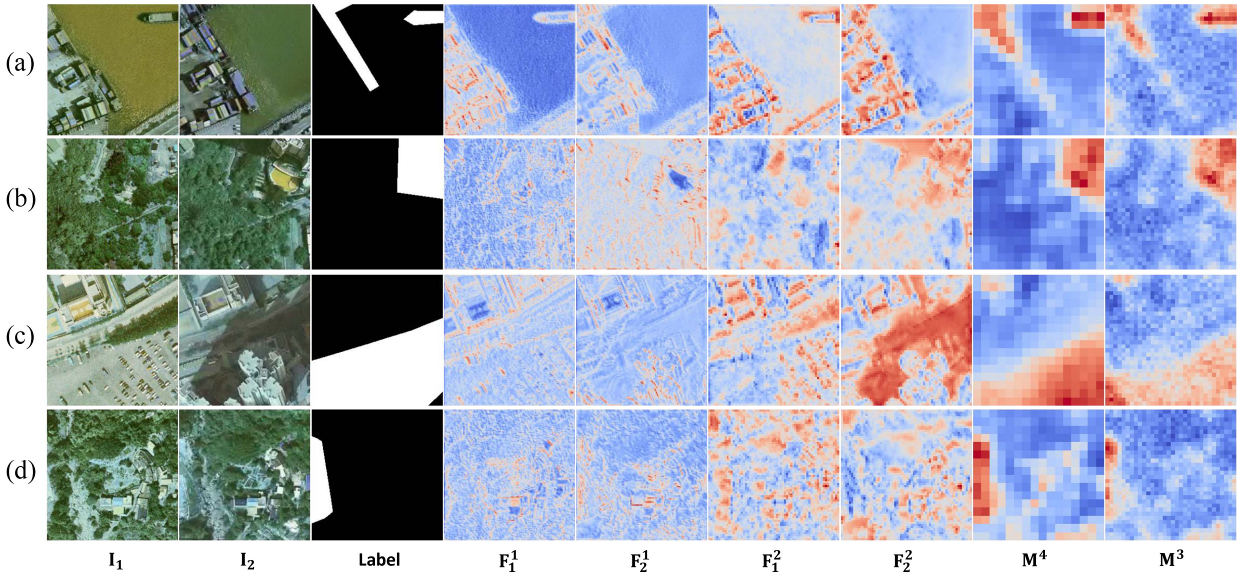

Fig. 6. Attention heat map visualization on sample pairs in the SYSU-CD dataset. The intensity of the color red indicates higher attention, while the intensity of blue signifies lower attention. This heat map demonstrates how effectively the model focuses on crucial regions within the image, highlighting the robustness of the attention mechanism in capturing changes in information. Selected samples are represented in figures (a) through (d).

combination is the most impactful. Removing this combination led to a decrease in the F1 score by 6.72 and 5.02 on the LEVIR-CD and SYSU-CD datasets, respectively.

E. Attention Heat Map Analysis

To delve deeper into the proposed model's mechanism, the semantic attention maps are visualized and analyzed based on the results produced on the two datasets. Fig. 6 illustrates the attention feature maps on the SYSU dataset. Evidently, the shallow layer features such as $\mathbf{F}_1^1$ and $\mathbf{F}_2^1$ primarily concentrate on local patterns and granular details, including edges and textures. As the depth of the layers increases, they start to identify global patterns and be aware of semantic information. In Fig. 6(b), the results are notably intriguing. The heat maps extracted from the encoder appear relatively faint and dispersed, suggesting that the network has yet to learn to attend on the building in the top-right corner. However, after passing through the 3th stage and 4th stage biattention layers, a substantial shift in focus is apparent. The attention is now primarily directed towards the top-right corner of the image, where the building is located. As per Fig. 6(c) and the $\mathbf{F}_2^2$ map, shadows considerably influence attention, with the heat mainly concentrated in shadowy areas. Nonetheless, after traversing the biattention mechanism, shadows cease to be an issue, and the attention reverts to regions with change hints. Our biattention is not flawless, as seen in Fig. 6(d). The $\mathbf{M}^4$ map is closer to the ground truth, but an additional stage of biattention resulting in the $\mathbf{M}^3$ map leads to a reduction in the attention region. Despite this, we tend to agree visually with the results derived from the SBA-PN network rather than the ground truth labels.

Likewise, as depicted in Fig. 7, shallow layers are tasked with learning local patterns, such as edges and textures, while deeper layers focus on context information. By comparing $\mathbf{M}^4$ and $\mathbf{M}^3$, it becomes clear that an additional biattention layer enables the network to learn more refined results. We wish to highlight the unique aspect of the LEVIR-CD dataset, which solely concentrates on building changes. We conjecture that this characteristic prompts our network to prioritize the building regions. For instance, in Fig. 7(b), upon examining $\mathbf{F}_1^2$ and $\mathbf{F}_2^2$, it is clear that attention is dominantly distributed throughout $\mathbf{F}_2^2$, meanwhile $\mathbf{F}_1^2$ exhibits minimal heat. The input images are nearly identical, with the only distinction being the presence of buildings in $\mathbf{I}_2$. Although the LEVIR-CD dataset is exclusively concerned with building changes, other alterations can affect our network's performance. In 7(c), we observe that both $\mathbf{M}^4$ and $\mathbf{M}^3$ allocate attention to not only building objects, but also road objects. In Fig. 7(d), despite the ground truth label indicating no changes across the entire region, a comparison of the images intuitively suggests that we are more inclined to agree with the network's learned M^3. The vivid red spots indeed seem to indicate building changes.

Fig. 8 offers a visualization of the ablation study concerning contrastive learning. From this figure, it becomes evident that the incorporation of contrastive supervision significantly improves the focus of the model on key objects within the images taken at two different time points. This is demonstrated by the heightened attention directed towards these crucial elements. In terms of the prediction feature, the presence of contrastive supervision appears to reduce the amount of irrelevant attention noise, thereby enhancing the overall clarity and relevance of the model's focus. This is particularly clear in Fig. 8(b), where the resultant feature map is noticeably refined. In the absence of contrastive supervision, the model seems to get distracted by the background elements, as indicated by a misplaced focus. Conversely, when contrastive supervision is applied, the attention appropriately shifts towards areas of the image where changes have occurred. This demonstrates the benefit of contrastive supervision in ensuring the model's attention is

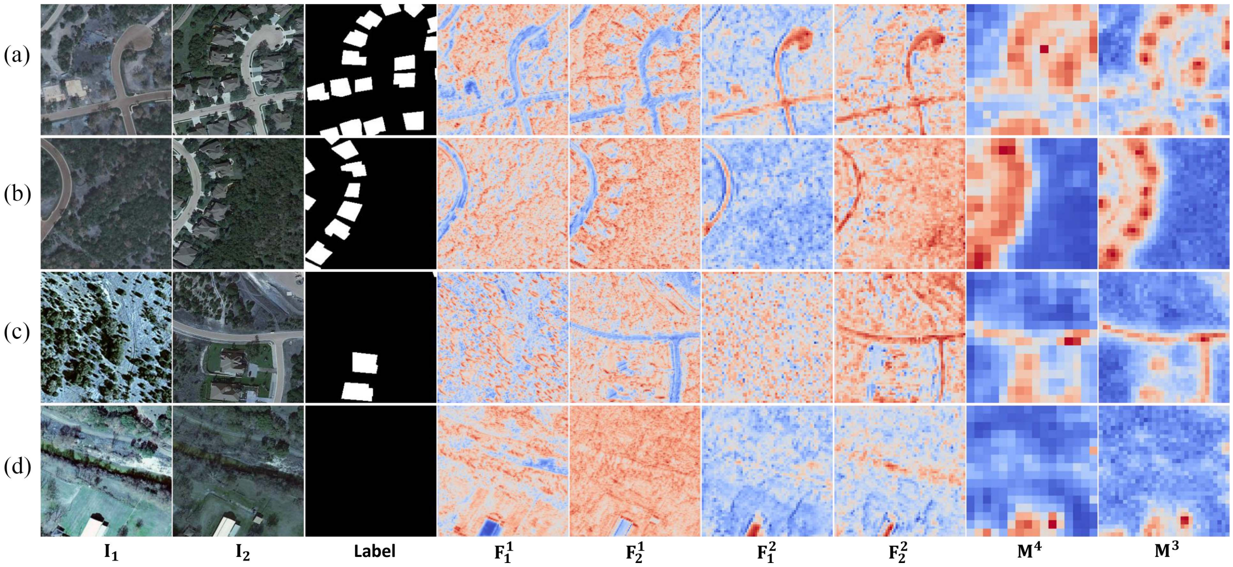

Fig. 7. Attention heat map visualization on sample pairs in the LEVIR-CD dataset. The color gradient, spanning from red to blue, illuminates the model's varying degrees of attention; red indicates high attention, whereas blue signifies low attention. This visualization underscores the effectiveness of the attention mechanism in highlighting and emphasizing significant change-related features within the image pairs of the LEVIR-CD dataset. Figures (a) through (d) represent selected samples.

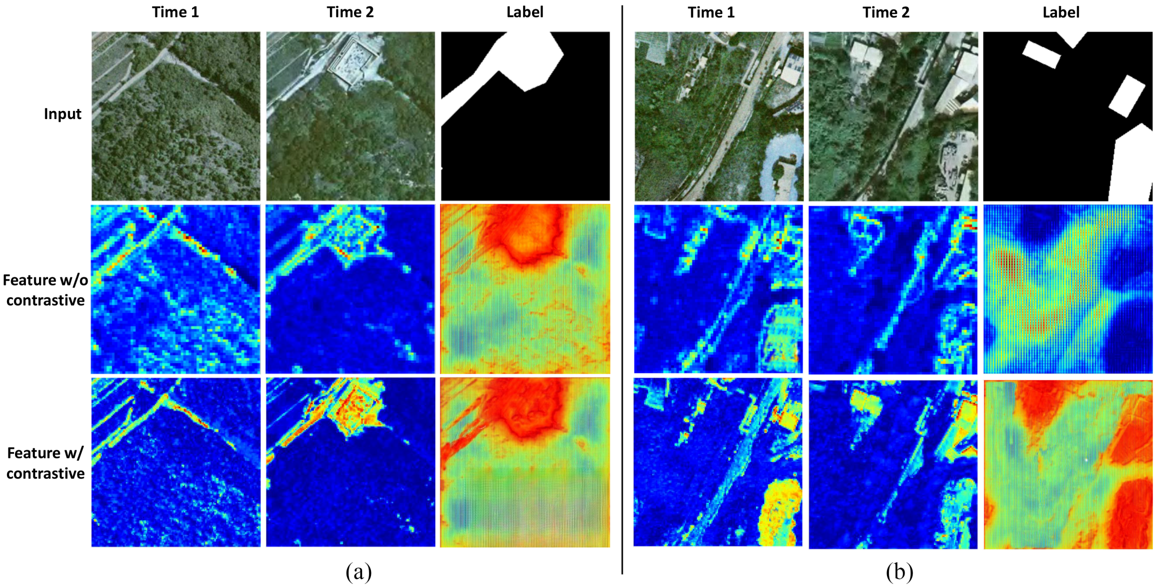

Fig. 8. Attention heat map visualization for contrastive ablation study on the SYSU-CD dataset. The color ranges from blue, which represents the lowest values, to red, which represents the highest values. In between, it transitions from blue to cyan, then to yellow, and finally to red as the value increases.

appropriately directed and meaningful patterns are more effectively recognized.

V. Conclusion

In this article, a novel DL architecture is presented, namely SBA-PN, to explore the broad-scale spatial information from both intrapair and interpair image patterns for change detection in remote sensing. The overall structure of SBA-PN follows a U-Net-like encoder-decoder structure. Specifically, a Siamese Transformer-like encoder formulates paired feature maps in a multiscale manner and a biattention-based decoder corresponding to this multiscale design formulates difference maps. Two pooling mechanisms are devised to emphasize the change relevant patterns, including a spatial optimal pooling module and a channel deviation pooling module. In addition, a contrastive pixel-wise supervision is devised for shallow encoder layers in pursuit of change-aware feature maps. Comprehensive experimental results on two widely used CD benchmark datasets demonstrate our proposed method is able to achieve the state-of-the-art performance.

In our future work, we aim to focus on improving semantic change detection with the SBA-PN. Our goal is to make the model better at identifying meaningful changes in various settings, including cities, natural areas, and during emergencies. Adding semantic analysis to our model is expected to increase its accuracy and usefulness in different situations. Also, exploring how hyperspectral bands affect change detection accuracy is important. Imaging data often have issues like degradation, noise, and variability [47]. Dealing with these issues is key to effectively modelling with hyperspectral imagery. We are working to advance change detection technology in remote sensing, making it stronger and more flexible for various needs.

References

[1] 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.

[2] D. Tomowski, M. Ehlers, and S. Klonus, "Colour and texture based change detection for urban disaster analysis," in *Proc. Joint Urban Remote Sens. Event*, 2011, pp. 329–332.

[3] 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.

[4] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 871.

[5] 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.

[6] H. Chen, X. Wu, S. Zeng, and Z. Wang, "Multi-scale attention based transformer u-net for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 1067–1070.

[7] P. R. Coppin and M. E. Bauer, "Digital change detection in forest ecosystems with remote sensing imagery," *Remote Sens. Rev.*, vol. 13, no. 3-4, pp. 207–234, 1996.

[8] N. A. QUARMBY and J. L. CUSHNIE, "Monitoring urban land cover changes at the urban fringe from spot HRV imagery in south-east England," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 953–963, 1989.

[9] A. K. Ludeke, R. C. Maggio, and L. M. Reid, "An analysis of anthropogenic deforestation using logistic regression and GIS," *J. Environ. Manage.*, vol. 31, no. 3, pp. 247–259, 1990.

[10] Y. Bayarjargal, A. Karnieli, M. Bayasgalan, S. Khudulmur, C. Gandush, and C. Tucker, "A comparative study of NOAA–AVHRR derived drought indices using change vector analysis," *Remote Sens. Environ.*, vol. 105, no. 1, pp. 9–22, 2006.

[11] K. N. C. author, K. Vaesen, B. Muys, and P. Coppin, "Comparative performance of a modified change vector analysis in forest change detection," *Int. J. Remote Sens.*, vol. 26, no. 5, pp. 839–852, 2005.

[12] J. S. Deng, K. Wang, Y. H. Deng, and G. J. 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.

[13] 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.

[14] 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.

[15] 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.

[16] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 484.

[17] 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.

[18] 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.

[19] 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.

[20] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[21] F. I. Diakogiannis, F. Waldner, and P. Caccetta, "Looking for change? roll the dice and demand attention," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3707.

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

[23] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1243–1251.

[24] 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.

[25] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Review articledigital change detection methods in ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[26] O. Yousif and Y. Ban, "Improving urban change detection from multitemporal SAR images using PCA-NLM," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2032–2041, Apr. 2013.

[27] H. Zheng et al., "Bi-CCD: Improved continuous change detection by combining forward and reverse change detection procedure," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8016605.

[28] Z. Du, X. Li, J. Miao, Y. Huang, H. Shen, and L. Zhang, "Concatenated deep-learning framework for multitask change detection of optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 719–731, 2024.

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

[30] 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.

[31] 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.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, vol. 30, 2017.

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

[34] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

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

[36] W. Yu et al., "Metaformer is actually what you need for vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10819–10829.

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

[38] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[39] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2015, vol. 38, pp. 562–570.

[40] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet : Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[41] L. Zhang, X. Chen, J. Zhang, R. Dong, and K. Ma, "Contrastive deep supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–19.

[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[43] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[44] 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.

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

[46] J. Zhang et al., "AERNet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617116.

[47] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

Hengzhi Chen received the Bachelor of Computer Science degree in data science from the University of Queensland, Brisbane, Australia, in 2021, and the Bachelor of Computer Science degree with Honours Class I, in 2022, from The University of Sydney, Sydney, NSW, Australia, where he is currently working toward the Ph.D. degree in change detection in remote sensing images, utilizing deep learning approaches to identify and analyze temporal changes in the earth's surface captured through remote sensing imagery with the School of Computer Science.

His research interests include computer vision, remote sensing, deep learning, and large language models.

Kun Hu received the B.Sc. degree in information and computing science and B.Eng. degree in software engineering from Shandong University, Jinan, China, in 2013, the M.Sc. degree in applied mathematics from Shanghai Jiao Tong University, Shanghai, China, in 2016, and the Ph.D. degree in information technology from The University of Sydney, Sydney, NSW, Australia, in 2022.

He is currently a Postdoctoral Research Associate with the School of Computer Science, The University of Sydney. His research interests include multimedia computing, computer vision, and graphics.

Patrick Filippi received the B.Sc. degree in agriculture (Hons I) from The University of Sydney, in 2013, with a major in agronomy and received the Ph.D. degree in agricultural science from The University of Sydney, in 2017.

His work was focussed on modelling and mapping spatio-temporal changes in agronomically-important soil properties across a semi-arid irrigated cotton-growing region of Australia. He is a Lecturer in precision crop management with the School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Sydney, NSW, Australia. His research in precision crop management is centred on using diverse on-farm and off-farm spatial and temporal datasets and data analytics to model and understand the variation in crop condition, yield, and quality. A strong component of this research is also on assessing how soil, weather, and management factors impact this variability, which can then guide improved management decisions going forward. This research has positive outcomes for farmers, agronomists, and a range of other stakeholders in terms of efficiency, profitability, and environmental sustainability.

Dr. Filippi is a Member of USyd's Precision Agriculture Laboratory and the Sydney Institute of Agriculture.

Wei Xiang (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees, both in electronic engineering, from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively, and the Ph.D. degree in telecommunications engineering from the University of South Australia, Adelaide, Australia, in 2004.

He is Cisco Research Chair of AI and IoT, Executive Director of the Victorian Centre for AI and Medical Innovation, and Director of the Cisco-La Trobe Centre for AI and IoT, La Trobe University, Melbourne, VIC, Australia. He was Foundation Chair and Head of Discipline of IoT Engineering with James Cook University, Cairns, Australia. He has authored or coauthored more than 450 peer-reviewed papers including three books and nearly 300 journal articles. His research interests include the Internet of Things, wireless communications, machine learning for IoT data analytics, and computer vision.

Dr. Xiang was the recipient of the TNQ Innovation Award in 2016, and Pearcey Entrepreneurship Award in 2017, and Engineers Australia Cairns Engineer of the Year in 2017, and corecipient of four Best Paper Awards at WiSATS'2019, WCSP'2015, IEEE WCNC'2011, and ICWMC'2009. Due to his instrumental leadership in establishing Australia's first accredited Internet of Things Engineering degree program, he was inducted into Pearcey Foundation's Hall of Fame in 2018. He is a TEDx speaker and an elected Fellow of the IET in U.K. and Engineers Australia. He has been awarded several prestigious fellowship titles. He was named a Queensland International Fellow (2010–2011) by the Queensland Government of Australia, an Endeavour Research Fellow (2012–2013) by the Commonwealth Government of Australia, a Smart Futures Fellow (2012–2015) by the Queensland Government of Australia, and a JSPS Invitational Fellow jointly by the Australian Academy of Science and Japanese Society for Promotion of Science (2014–2015). He was the Vice Chair of the IEEE Northern Australia Section from 2016 to 2020. He is currently an Associate Editor for *IEEE Communications Surveys & Tutorials*, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *IEEE Internet of Things Journal*, *IEEE Access*, and Nature journal of *Scientific Reports*. He has served in a large number of international conferences in the capacity of General Co-Chair, TPC Co-Chair, Symposium Chair, etc.

Thomas Bishop received the Ph.D. degree in precision agriculture from The University of Sydney, Sydney, NSW, Australia, in 2002.

His main teaching is related to applied statistics, environmental science, and GIS. He is the Academic Director of the Sydney Informatics Hub which is a Core Research Facility, The University of Sydney. He is a Professor with the School of Life and Environmental Science, The University of Sydney. Prior to starting work at the University of Sydney in 2007, he held postdoctoral positions with the University of Florida, FL, USA, Rothamsted Research, Harpenden, U.K., and the University of New South Wales, Sydney, NSW, Australia. His research interests include modeling and predicting the variation of environmental properties in space and time with an emphasis on applying this to the domains of soil, agriculture and hydrology.

Dr. Bishop is an Associate Editor for the *European Journal of Soil Science and Soil Research* and on the Editorial Board of *Geoderma* and *Pedosphere*.

Zhiyong Wang (Member, IEEE) received the B.Eng. and M. Eng. degrees in electronic engineering from the South China University of Technology, Guangzhou, China, in 1996 and 1999, respectively, and the Ph.D. degree from Hong Kong Polytechnic University, Hong Kong, in 2003.

He is a Professor and Director of the Multimedia Computing Laboratory, School of Computer Science, The University of Sydney, Sydney, NSW, Australia. His research interests include multimedia computing and its applications in agriculture, earth observation, health, and medicine, including multimedia information retrieval, summarization, multimedia data mining, multimedia content understanding (e.g., human action recognition and affective analysis), multimedia content creation, human-centered multimedia computing, computer vision, computer graphics, remote sensing, and pattern recognition.