# 3D Neighborhood Cross Differencing: A New Paradigm Serves Remote Sensing Change Detection

Wei Jing, Kaichen Chi, Qiang Li, *Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

*Abstract*—**Change detection is a prevalent technique in remote sensing image analysis for investigating geomorphological evolution. The modeling and analysis of difference features are crucial to for the precise detection of land cover changes. In order to extract difference features, previous work has either directly computed them through differential operations or implicitly modeled them via feature fusion. However, these rudimentary strategies rely heavily on a high degree of congruence within the bi-temporal feature space, which results in the model's diminished capacity to capture subtle variations induced by factors such as differences in illumination. In response to this challenge, the concept of 3D Neighborhood Difference Convolution (3D-NDC) is proposed for robustly aggregating the intensity and gradient information of features. Furthermore, to delve into the deep disparities within bi-temporal instance features, we propose a novel paradigm for differential feature extraction based on 3D-NDC, termed as 3D Neighborhood Cross Differencing. This strategy is dedicated to exploring the interplay of cross-temporal features, thereby unveiling the inherent disparities among various land cover characteristics. Additionally, a Detail-focused Refinement (DfR) decode based on the Laplace operator has been designed to synergize with the 3D Neighborhood Cross Differencing, aiming to improve the detail performance of change instances. This integration forms the basis of a new change detection framework, named ChangeLN. Extensive experiments demonstrate that ChangeLN significantly outperforms other state-of-the-art change detection methods. Moreover, the 3D Neighborhood Cross Difference strategy exhibits the potential for integration into other change detection frameworks to improve detection performance. Open code is available from https://github.com/weiAI1996/3DNCD_ChangeLN.**

*Index Terms*—**Remote sensing image, Change detection, Deep learning, 3D Neighborhood difference convolution, Cross differencing, Detail-focused refinement.**

## I. Introduction

CHANGE detection is the process of observing land cover evolution by interpreting multi-temporal images [1]–[4]. With the vigorous development of earth observation technologies, it has greatly facilitated the implementation of tasks such as disaster assessment [5], agricultural investigation [6], and urban planning [7].

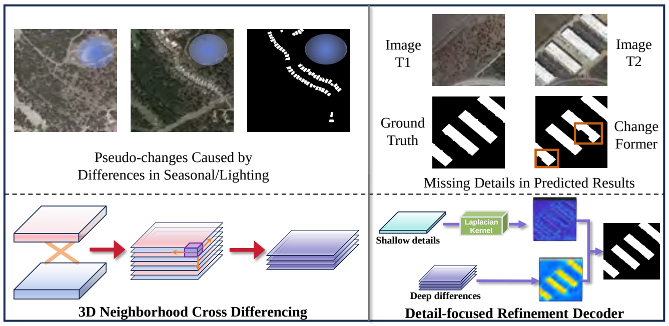
Pseudo-changes Caused by
Differences in Seasonal/Lighting
Image T1
Image T2
Ground Truth
Change Former
Missing Details in Predicted Results
3D Neighborhood Cross Differencing
Shallow details
Laplacian Kernel
Deep differences
Detail-focused Refinement Decoder

Fig. 1. Challenges in Change Detection and Proposed Solutions.

To enhance the detection of surface changes from remote sensing imagery, traditional change detection methods predominantly focus on extracting change information based on pixel spectral differences. These include techniques such as differencing, ratioing, and regression analysis [8], [9]. To further capitalize on the spectral information available in remote sensing images, researchers have incorporated methods like Change Vector Analysis [10], Principal Component Analysis [11], and Slow Feature Analysis [12] into change detection tasks. However, the advent of Very High Resolution (VHR) remote sensing imagery, while offering more detailed information, also presents significant challenges [13], [14]. Firstly, high-resolution imagery increases data processing requirements, placing greater demands on storage and computational capacity. Secondly, the information density in VHR imagery is significantly higher, potentially containing more noise and interference. This necessitates the development of more refined and robust interpretation algorithms to effectively process remote sensing data. Consequently, these methods are gradually becoming less prevalent in the analysis of VHR imagery due to their diminished effectiveness in handling the complexities associated with such high-resolution data.

Recently, deep learning has demonstrated powerful representation learning capabilities to extract deep semantic information inherent in images [15]–[17]. Supported by several widely utilized benchmarks [18]–[21], deep learning-based frameworks for change detection have flourished and can generally be categorized into two types. The first type involves early fusion of bi-temporal images followed by the application of segmentation algorithms to detect areas of change [22]. The second type maps image pairs into a latent feature space and then extracts their differential features for classification purposes [23]. Compared to early fusion strategies, the latter approach, benefiting from enhanced resistance to environmen-

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2041; and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University, China, under Grant CX2024108. (*Corresponding author: Qi Wang.*)

Wei Jing is with the National Elite Institute of Engineering, and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: wei_adam@mail.nwpu.edu.cn).

Kaichen Chi, Qiang Li and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (chikaichen@mail.nwpu.edu.cn, liqmges@gmail.com, crabwq@gmail.com).

tal interference, is gradually becoming the dominant strategy for change detection in VHR imagery.

After significant development over several years, change detection based on deep learning has reached a mature stage. Nonetheless, some unresolved issues remain. (1) The two most commonly used differential feature extraction operators, point-to-point differencing, and channel-level cascading, focus solely on spatial characteristics and overlook the modeling of channel relationships in concatenated bi-temporal features. Consequently, they are highly vulnerable to pseudo-changes, such as seasonal and lighting differences, posing challenges in ensuring robustness against non-semantic changes. (2) Continuous downsampling in the feature extraction process leads to the loss of detailed information, resulting in the partial absence of change instances.

To address the aforementioned issues, we develop two targeted and feasible solutions, as illustrated in Fig. 1. Firstly, inspired by difference convolution [24], a novel convolution operator is proposed, called the 3D Neighborhood Difference Convolution (3D-NDC). The 3D-NDC operator excels in depicting fine-grained spatiotemporal invariant information. Furthermore, for robust extraction of bi-temporal differential features, we develop the 3D Neighborhood Cross Difference strategy based on 3D-NDC. Additionally, to mitigate the loss of detail in instances of change, we design a Detail-focused Refinement (DfR) decoder by integrating low-level detail information with the second-order Laplacian operator, maintaining the boundaries of the instances. The main contributions of this work can be summarized as follows:

1. We developed the ChangeLN network for very high resolution (VHR) image change detection. This network employs a 3D Neighborhood Cross Difference strategy to model difference features and a well-designed decoder to enhance detail performance. Extensive experiments demonstrate that ChangeLN significantly outperforms other methods in both quantitative and qualitative evaluation.

2. A novel convolutional operator, 3D-NDC, is proposed, providing robust fine-grained representation way of bi-temporal twin features. It effectively replaces conventional convolutions in siamese networks for modeling explicit difference features.

3. To robustly model differential information, we develop a 3D Neighborhood Cross Difference strategy based on 3D Neighborhood Difference Convolution. This module analyzes spatiotemporal differences in bi-temporal geospatial data from both spatial and channel dimensions. This approach offers a new paradigm for differential feature extraction, which significantly enhances performance when embedded in similar network architectures, with minimal additional overhead.

4. To improve the detail accuracy in detecting changes, we design a Detail-focused Refinement decoder. The decoder reinforces the semantic boundaries of change instances by explicitly extracting higher-order gradients from low-level features.

## II. Related work

### A. Deep Learning-Based Change Detection

Thanks to the commercialization of high-resolution and submeter level remote sensing satellites, significant advancements have been made in remote sensing image change detection [25], [26]. Initially, early change detection algorithms executed image algebra operations on a per-pixel basis [8]. Subsequently, these methods evolved to incorporate machine learning techniques, with Support Vector Machines (SVMs) being a notable example [10]. However, these approaches have become inadequate in keeping up with the ever-increasing spatial and spectral resolution of images. In comparison to traditional change detection algorithms, deep learning offers a more robust feature representation capability [27], [28]. It can simultaneously capture both low-level details and high-level semantics in images, making it suitable for a variety of complex scenarios. Early and late fusion are representative strategies of change detection based on deep learning, adept at adapting to diverse and intricate environments.

Early fusion strategies typically involve the fusion of images from two different time points at the input stage of a network using differential or concatenation techniques, and then consider change detection as a panoramic segmentation problem [22], [29]. Sun et al. [22] integrated an extended ConvLSTM structure into the U-Net framework to achieve end-to-end change detection and further utilized Atrous convolution to mine multi-scale spatial information. Lin et al. [29] analogized change detection to video understanding tasks, explicitly considering the spatiotemporal coupling issues within the encoder. They proposed the P2V-CD model to decouple the spatiotemporal dimensions of paired temporal images.

Siamese networks, as a representative model of post-fusion strategies, aim to map bi-temporal images into a closely related feature space in order to extract differential features for change detection purposes [19], [30], [31]. Zhang et al. [19] aimed to improve the boundary coherence and internal cohesion of objects in the resultant change maps by developing a deeply supervised difference discrimination network. This network enhances the change map by fusing deep features from original inputs with difference features of bi-temporal images. Lei et al. [31] addressed the challenge of identifying irrelevant changes. They utilized a siamese network to distinguish between foreground and background, thereby obtaining a discrepancy representation. Additionally, they incorporated remote relationships to enhance the edge coherence and internal cohesion of the changing objects. Recently, transformer models [32], [33] have shown exceptional capabilities in the realm of computer vision, outperforming CNN-based methods in various tasks, including classification and detection. The core self-attention mechanism enables the model to capture the correlation between different regions of the image, thus modeling global dependencies [34]–[37]. Bandara et al. [38] unified the hierarchical Transformer structure in siamese networks, effectively rendering the multi-scale remote details required for precise change detection, achieving SOTA performance on multiple CD datasets.

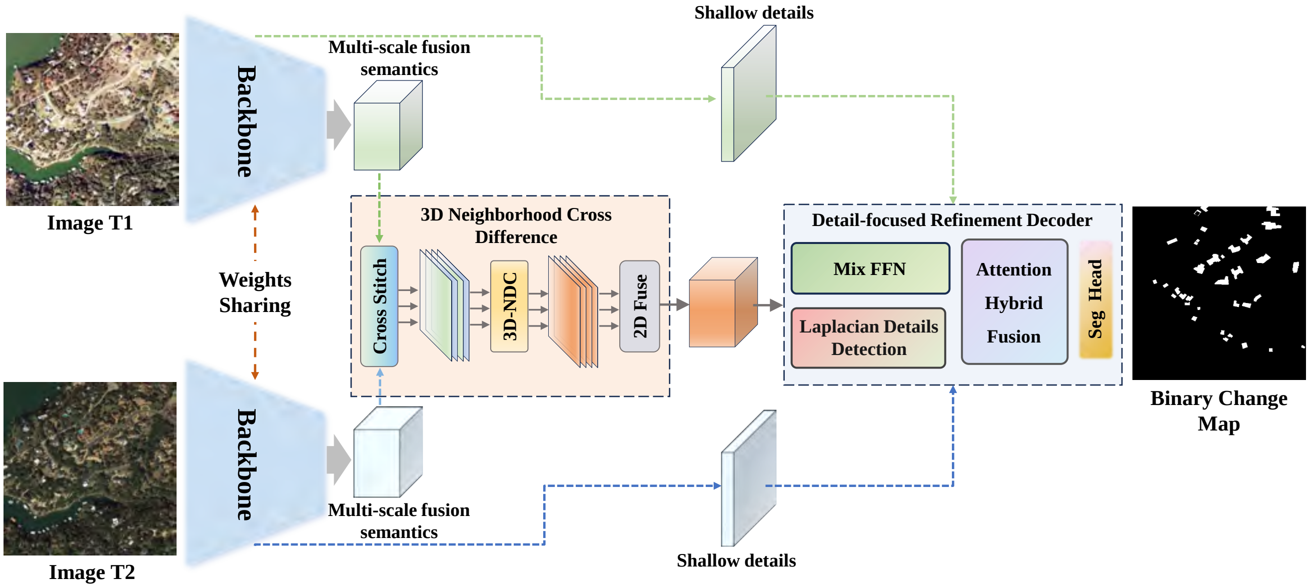

Fig. 2. Overall Framework of ChangeLN. The 3D Neighborhood Cross Difference module is used to model the bi-temporal difference features and the Detial-focused Refinement Decoder is used to predict the change map while preserving details.

### B. Difference Feature Modeling and Enhancement

The extraction of discriminative differential features is a prerequisite for achieving high-precision change detection. The strategy of modeling through convolutional layers after differencing and concatenation is widely employed in DL-based change detection methodologies. Daudt et al. [23] designed three distinct frameworks based on differencing and concatenation operations and comprehensively compared the performance of various strategies.

Recent research has been focused on enhancing the extraction of differential features through strategies such as attention mechanisms and multi-scale feature fusion. Guo et al. [39] introduced an iterative differential-enhanced transformer to optimize differential features, emphasizing changes and suppressing unchanged regions. Furthermore, they proposed a hierarchical fusion strategy to combine differential features at multiple scales. Lv et al. [40] introduced spatial-spectral attention mechanism and multi-scale dilated convolution module in an attempt to capture more salient changes for further enhancing detection accuracy. Luo et al. [41] proposed a multiscale diff-changed feature fusion network (MSDFFN) to enhance feature representation by learning refined change components between bi-temporal hyperspectral images across different scales. A diff-feature contrast enhancement network (DCENet) [42] has also been developed to leverage a limited number of labeled samples and a large number of unlabeled samples to enhance detection confidence in semi-supervised hyperspectral image CD. To enhance the network's ability to discriminate change features, Zhao et al. [43] employed a dual-channel attention module to fuse features extracted from both dual-stream and single-stream encoders. While the aforementioned researchers extracted and enhanced difference features from different perspectives, these methods still did not break free from feature concatenation and implicit modeling through 2D convolutions.

In order to capture robust discriminative differential features, we propose a novel differential strategy, termed 3D Neighborhood Cross Differencing, to explore the spatiotemporal variations of bi-temporal features in both spatial and channel dimensions. This approach aims to improve the modeling of high discriminative differential features.

## III. METHODOLOGY

In this section, we first introduce the change detection framework ChangeLN, which is constructed based on the 3D Neighborhood Cross Difference strategy and Detail-focused Refinement Decoder. Then, two key components of ChangeLN are described in detail.

### A. Overall Framework of ChangeLN

To capture the subtle inherent differences in bi-temporal images and preserve the edge details of changing instances, an end-to-end change detection network called ChangeLN is proposed. This network comprises a 3D neighborhood cross difference strategy and a detail-focused refinement decoder.

As shown in Fig. 2, ChangeLN takes bi-temporal images, T1 and T2, as inputs and utilizes a Siamese backbone to extract spatiotemporal features at different hierarchical levels. These spatiotemporal features are fused through the concatenation of channel dimensions and dimensionality reduction operations via linear convolution, as follows:

$$X_s = \mathcal{K}_{1\times1}\left(\Pi_{i=1}^{n} X_i\right), \tag{1}$$

where $\mathcal{K}_{1\times1}$ is linear convolution kernel, $\Pi$ denotes the operation of concatenation, $n$ represents the number of feature hierarchy levels in the backbone output. Using the proposed

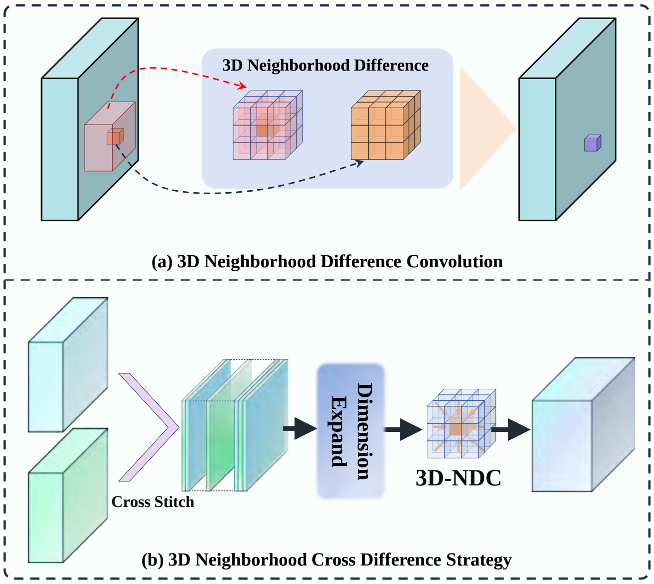
3D Neighborhood Difference
(a) 3D Neighborhood Difference Convolution
Cross Stitch
Dimension Expand
3D-NDC
(b) 3D Neighborhood Cross Difference Strategy

Fig. 3. Structure of the 3D Neighborhood Difference Convolution.

3D neighborhood cross difference strategy, we explicitly extract inherent difference information. The computation is as follows:

$$X_{diff} = \mathcal{F}_{fuse}\left(X_s^1 \Theta X_s^2\right), \tag{2}$$

where $X_s^1$ and $X_s^2$ are the multi-scale fusion semantics of the two temporal phases, $\Theta$ denotes 3D neighborhood cross differencing, and $F_{fuse}$ denotes the 2D convolution-based feature fusion process. $F_{fuse}$ is calculated as follows:

$$F_{fuse} = \mathcal{G}\left(\mathcal{K}_{3\times3}\left(X\right)\right), \tag{3}$$

where $\mathcal{G}$ denotes GELU activate function. Shallow features extracted by the backbone typically contain an abundance of detailed information, which is beneficial for preserving the complete boundaries of changing instances. In ChangeLN, the Detail-focused Refinement Decoder is proposed to learn the critical details within the shallow-level features and fused them with the bi-temporal deep-level differences. Finally, a straightforward segmentation head is employed to predict a binary change map. The cross-entropy loss is adopted to supervise the predicted change maps, which can be expressed as:

$$\mathcal{L}_{cd} = -\sum_{c=1}^{M}\left(y_c \log\left(p_c\right)\right), \tag{4}$$

where $M$ indicates the number of categories, $c$ is the category index, $y_c$ represents the probability of the true class being $c$, and $p_c$ represents the probability of the predicted class being $c$.

### *B. 3D Neighborhood Cross Differencing*

The robust modeling of difference information is a prerequisite for high-precision change detection. However, due to the imaging condition disparities in bi-temporal images, the phenomenon of ”same-object-different-spectrum” and ”same-spectrum-different-object” severely interferes with the extraction of change features. Furthermore, the change features extracted using existing point-to-point differencing and 2D convolution implicit modeling strategies also struggle to ensure the model’s generalization performance.

Inspired by the Local Binary Pattern (LBP) [24], we introduce neighborhood differencing into 3D convolution to mitigate pseudo-changes caused by imaging condition disparities. 3D spatial convolution is a common operation for modeling temporal visual features and consists of two main steps: 1) Sampling the three-dimensional local receptive field region $\mathcal{R}$ on the input feature map $X^{B\times C\times T\times H\times W}$, where $B$ represents the batch size, $C$ represents the number of channels, $T$ represents the temporal or depth dimension, $H$ represents the height and $W$ represents the width. 2) Aggregating the sampled values through weighted summation. Therefore, the output feature map $Y$ can be formulated as follows:

$$Y\left(p_0\right) = \sum_{p_n\in\mathcal{R}} W\left(p_n\right)\cdot X\left(p_0+p_n\right), \tag{5}$$

where $p_0$ denotes the current position on the input and output feature maps, and $p_n$ enumerates the positions in $\mathcal{R}$.

Similarly, 3D Neighborhood Difference convolution also comprises two steps, namely sampling and aggregation. The sampling step is akin to that in 3D convolution, but the aggregation step differs. As depicted in Fig. 3(a), 3D-NDC tends to aggregate neighborhood difference information from sampled points in the three-dimensional space, and can be expressed as follows:

$$Y\left(p_0\right) = \sum_{p_n\in\mathcal{R}} W\left(p_n\right)\cdot\left(X\left(p_0+p_n\right) - X\left(p_0\right)\right). \tag{6}$$

By computing the difference information of adjacent pixels, we aim for 3D-NDC to capture discriminative features robust to factors such as illumination and seasonality.

In order to explicitly extract the difference information from the bi-temporal features, we propose a 3D Neighborhood Cross difference strategy. As shown in Fig. 3(b), bi-temporal features are no longer directly concatenated but are instead cross-stitched channel-wise as follows:

$$X_{cs} = \mathcal{F}_{cs}\left(X_1, X_2\right). \tag{7}$$

In cross-stitched feature $X_{cs}$, channels are alternately sourced from $X_1$ and $X_2$. Specifically, $X_1$ contributes channels at even indices (e.g., 0, 2, 4, 6...), while $X_2$ provides channels at odd indices (e.g., 1, 3, 5, 7...).

Further, we employ 3D-NDC to handle cross-stitched features. Unlike 2D convolution, which operates directly across the entire channel dimension, 3D-NDC computes local differentce features $X_{diff}$ by sliding across both spatial and channel dimensions. To further learn the intrinsic nature of the difference information, an additional 2D convolution is finally applied to $X_{diff}$.

### *C. Detail-focused Refinement Decoder*

To improve the detail performance of the change maps, we propose the Detail Focus Reinforcement decoder (DfR),

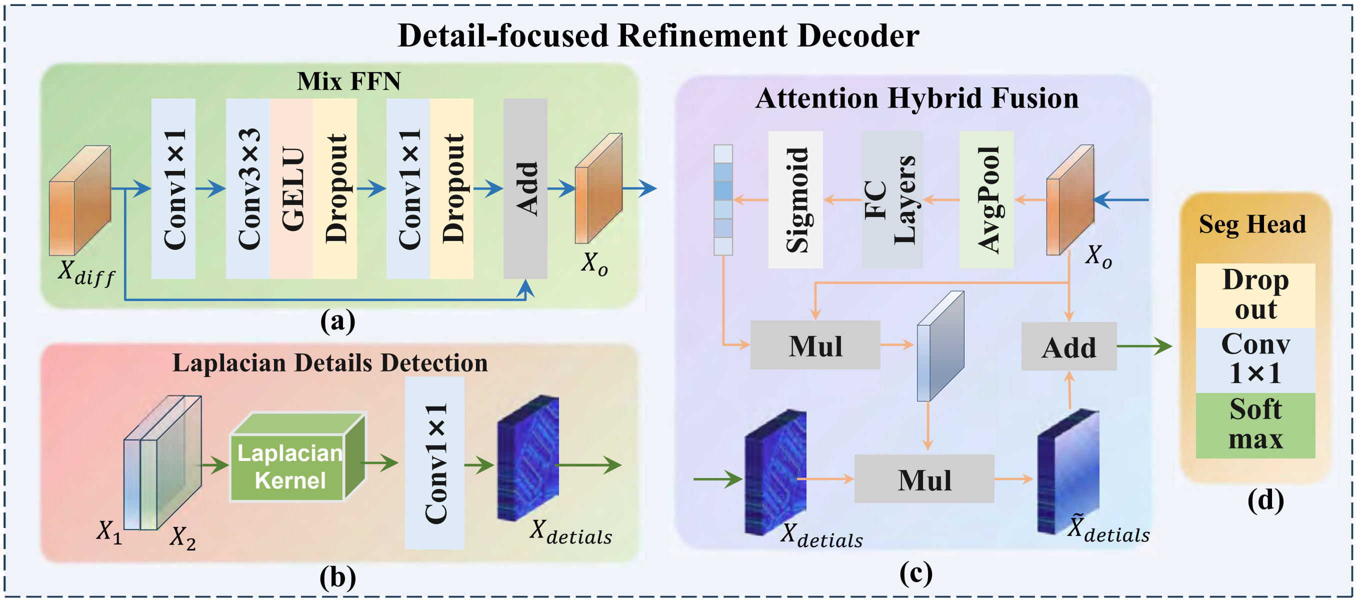

Fig. 4. Structure of the Detail-focused Refinement Decoder.

designed to preserve the integrity of change instances by supplementing details in the difference information. The DfR consists of four clearly defined sub-units: Mixed Feedforward Network (Mix FFN), Laplacian Details Detection, Attention Hybrid Fusion, and Seg Head, as illustrated in Fig. 4.

In Fig. 4(a), Mix FFN serves as the feedforward layer in the DfR architecture. While Vision Transformer (ViT) employs positional encoding (PE) to introduce position information, when the testing resolution differs from the training resolution, interpolation of positional encoding is required, often leading to a decrease in accuracy. Taking into account the impact of zero-padding on positional information leakage, we introduce the Mix FFN, which can be formulated as follows:

$$X_o = \mathcal{K}_{1\times1}\left(\mathcal{G}\left(\mathcal{K}_{3\times3}\left(\mathcal{K}_{1\times1}\left(X_{diff}\right)\right)\right)\right) + X_{diff}, \tag{8}$$

where $X_{diff}$ is the feature from 3D Neighborhood Cross Differencing module. The Laplacian operator is a second-order gradient operator that effectively detects edge information in an image by identifying extremum points (i.e., zero-crossings) of intensity change rates. To explicitly extract detailed information from shallow features, we construct the Laplacian Detail Detection unit with the Laplacian operator as its core, formulated as:

$$X_{detials} = \mathcal{K}_{1\times1}\left(\mathcal{K}^{lap}_{3\times3}\left(X_1 \parallel X_2\right)\right) \tag{9}$$

where $X_1$, $X_2$ are bi-temporal shallow features from the siamese network, $\parallel$ represents feature concatenation operation, and $\mathcal{K}^{lap}_{3\times3}$ denotes a Laplacian kernel of size $3 \times 3$. To effectively integrate detailed features with bi-temporal difference features, an Attention Hybrid Fusion unit is designed. Initially, global pooling and fully connected layers are employed to generate channel-specific semantic responses:

$$V = \mathcal{F}_{FC}(\mathcal{F}_{GP}(X_o)), \tag{10}$$

where $\mathcal{F}_{FC}$ refers to the fully connected layer, and $\mathcal{F}_{GP}$ denotes the operation of global pooling. The attention difference features are generated using channel multiplication as follows:

$$X_a = X_o \circ V, \tag{11}$$

where $\circ$ denotes the operation of channel-wise multiplication. Detail feature masking and smoothing is achieved by multiplying the attention difference feature with the detail feature as follows:

$$\tilde{X}_{detials} = X_{detials} \odot X_o, \tag{12}$$

where $\odot$ denotes the operation of point-to-point multiplication Finally, the feedforward features and smoothed details are added to output detail-focused difference features $X_{dfd}$. A simple segmentation head is used to predict the variation graph as follows:

$$y = \mathcal{S}\left(\mathcal{K}_{1\times1}\left(X_{dfd}\right)\right), \tag{13}$$

where $\mathcal{S}$ denotes softmax function.

## IV. EXPERIMENTS

In this section, to evaluate the performance of the proposed method, a set of experiments was carried out in this section using three openly accessible datasets: LEVIR-CD [18], DSIFN [19], and SVCD [20].

### A. Dataset Description and Experimental Setting

1) LEVIR-CD [18]: This dataset comprises 637 images with a very high resolution of 0.5 meters per pixel. Each image has dimensions of $1024 \times 1024$ pixels. These bi-temporal images exhibit significant land cover changes, particularly noticeable in the growth of buildings, over a time span ranging from 5 to 14 years. The dataset consists of 31,333 change instances in total. We use standardized training, validation, and testing sets. During model training, validation, and testing, we extracted

non-overlapping patches from the original images using the sliding window approach, with each patch having dimensions of $512 \times 512$.

2) DSIFN-CD [19]: This dataset is a publicly available collection of Change Detection data manually gathered from Google Earth and comprises six high-resolution bi-temporal images covering six cities in China, namely Beijing, Chengdu, Shenzhen, Chongqing, Wuhan, and Xi'an. Five of the images (corresponding to Beijing, Chengdu, Shenzhen, Chongqing, and Wuhan) were subdivided into 394 sub-image pairs, each sized at $512 \times 512$ pixels. Following data augmentation procedures, a total of 3940 bi-temporal image pairs were obtained. Additionally, the Xi'an image pairs were subdivided into 48 pairs specifically used for model testing.

3) SVCD [20]: This dataset comprises 16,000 pairs of authentic seasonal change remote sensing images obtained from Google Earth. Each pair of images is composed of pixels with dimensions of $256 \times 256$ and varying spatial resolutions ranging from 3 to 100 cm per pixel. The dataset is divided into training, validation, and testing sets, with 10,000, 3,000, and 3,000 samples, respectively. In contrast to the WHU-CD dataset, the SVCD dataset offers a larger sample size, enabling a comprehensive evaluation of the fitting capability. Moreover, since the majority of the bi-temporal image pairs originate from different seasons, this dataset serves as a valuable benchmark to assess the robustness of models against the same-object different-spectrum phenomena induced by seasonal factors.

### B. Evaluation Metrics

The broadly used criteria, Intersection over Union (IoU), Precision, Recall, and F1-score (F1), are applied for quantitative evaluation of change detection.

**IoU** is a metric defined as the ratio of the intersection of predicted and ground truth regions to their union. It is calculated as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{FP} + \mathrm{TP}}, \tag{14}$$

where TP, TN, FP, and FN refer to the true positive, true negative, false positive, and false negative pixels , respectively.

**Precision** refers to the probability of correctly predicted positive samples (TP) out of all samples predicted as positive (TP+FP):

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}}. \tag{15}$$

**Recall** is defined as the proportion of correctly predicted positive samples (TP) out of all true positive samples (TP+FN) as follows:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{16}$$

**F1** is the harmonic mean of precision and recall, which provides a comprehensive measure of overall performance:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{17}$$

### C. Comparison with Advanced Methods

To evaluate the superiority of ChangeLN, we compare it with 9 SOTA algorithms. The brief descriptions of the comparative algorithms are as follows:

1) FC-Siam-Diff [23]: A fully convolutional Siamese network based on U-Net architecture that computes the absolute difference of multi-scale bi-temporal features as change-related features.
2) FC-Siam-Conc [23]: A fully convolutional Siamese network based on U-Net architecture that extracts difference information from multi-scale bi-temporal features using convolutional layers.
3) STANet [18]: A Siamese network equipped with a spatiotemporal attention module and a pyramidal spatiotemporal attention module designed to explore spatiotemporal relationships for change detection.
4) HANet [44]: A discriminant Siamese hierarchical attention network that uses a lightweight self-attention mechanism to integrate multi-scale features and refine detailed features.
5) IFN [19]: Fusion of multi-level deep features and image difference features using attention mechanisms, improving the integrity of change map boundaries and the compactness within regions.
6) SNUNet [45]: A Siamese network based on NestedUNet that aggregates and refines features from multiple semantic levels, suppressing semantic gaps and localization errors to some extent.
7) ChangFormer [38]: Hierarchical transformer encoders are used to extract bi-temporal features, and a feature difference module is designed to compute feature discrepancies at different scales.
8) PCAM [46]: A Siamese network leverages difference information through three strategies, ”align”, ”perturb ”and ”decouple”, to predict semantic changes in a content-aware and content-agnostic manner..
9) Changer [47]: A Siamese change detection architecture that explores interactions between bi-temporal features in a feature extractor while interactively aligning features using a flow-based dual-alignment fusion module.

**Comparison Experiments on the LEVIR-CD:** The results in Fig. 5 demonstrate the qualitative performance of various algorithms on the LEVIR-CD dataset, highlighting the outstanding detection capabilities of ChangeLN. As depicted in Fig. 5, Row 1, the buildings within the rectangular bounding boxes exhibit distinct spectral characteristics due to drastic changes in lighting conditions. Compared to competitors, ChangeLN exhibits fewer false positives. The building in Fig. 5, Row 2, differs from the other building features in the entire dataset in that most of the models show a large number of misses, and the misses exhibited by our algorithm are due to insufficiently fine-grained labeling. The visualizations in the third Row illustrate that ChangeLN outperforms other methods in preserving fine-grained details.

The quantitative results presented in the Table I indicate that ChangeLN, utilizing ResNet and Mit-b0 as its primary backbones, respectively, achieved the optimal and suboptimal

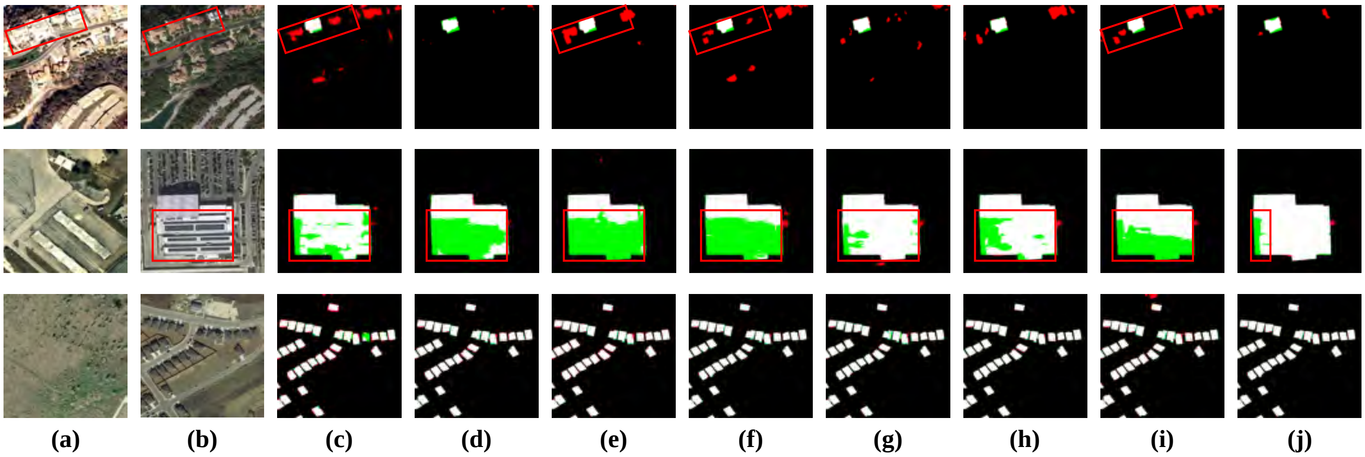

Fig. 5. The qualitative comparison of different methods on the LEVIR-CD dataset. Please zoom-in for the best view. (a) Pre-temporal image. (b) Post-temporal image. (c) STANet. (d) HANet. (e) IFN. (f) SNUNet. (g) ChangeFormer. (h) PCAM. (i) Changer. (j) Ours. We highlight the TP areas in white, the FP areas in red, and the FN areas in green. The black color denotes the TN areas.

TABLE I
QUANTITATIVE RESULTS ON THE LEVIR-CD. THE BEST RESULTS ARE MARKED IN BOLD, THE 2ND-BEST IS MARKED WITH UNDERLINE. * INDICATES RESNET BACKBONE REPLACED BY MIT-B0.

| Method | Backbone | #Param (M) | MACs (G) | Precision | Recall | F1 | IoU |
|---|---|---|---|---|---|---|---|
| FC-Siam-Diff | UNet | 1.35 | 17.06 | 89.85 | 80.42 | 84.88 | 73.72 |
| FC-Siam-Conc | UNet | 1.54 | 19.47 | 86.44 | 85.23 | 85.83 | 75.18 |
| STANet | ResNet18 | 12.21 | 50.21 | 85.26 | 88.39 | 86.80 | 76.68 |
| HANet | ResNet18 | 3.02 | 97.55 | 92.25 | 89.21 | 90.70 | 82.99 |
| IFN | VGG-16 | 35.99 | 316.52 | 90.97 | **91.13** | 91.05 | 83.57 |
| SNUNet | UNet++ | 12.03 | 46.92 | 92.53 | 89.85 | 91.17 | 83.77 |
| ChangeFormer | MiT-b1 | 13.94 | 26.42 | **93.35** | 89.28 | 91.22 | 83.94 |
| PCAM | ResNet18 | 111.00 | 85.30 | 92.50 | 90.36 | 91.41 | 84.18 |
| Changer | ResNet18 | 11.39 | 23.82 | 92.38 | 90.80 | 91.58 | 84.47 |
| ChangeLN | ResNet18 | 12.35 | 26.63 | 93.14 | 90.83 | **91.97** | **85.14** |
| ChangeLN* | MiT-b0 | 4.42 | 11.19 | 93.34 | 90.32 | 91.81 | 84.85 |

outcomes in terms of the comprehensive metrics F1 and IoU. Our methodology realized an F1 score of 91.97% and a mean IoU (mIoU) of 85.14%, surpassing Changer by 0.47 and 0.67 percentage points, respectively. This demonstrates the effectiveness of our approach in enhancing performance metrics in this domain.

**Comparison Experiments on the DSIFN:** Fig. 6 illustrates a visual comparison of different methods. Owing to the lower sample space resolution in the DSIFN dataset and the inherent reduction in feature resolution during the forward process of deep models, the sample instance boundaries in the change maps are relatively coarse. Compared to competitors, ChangeLN demonstrates superior performance in preserving boundaries. Moreover, the ground cover change from bare soil to playground, as highlighted in the rectangular frame in Row 2, is largely undetected by most algorithms. In Row 3, the complexity and density of the building distributions in the bi-temporal images, coupled with the spectral similarity of some buildings to impervious surfaces, result in partial omissions and false detections in comparative models. However, ChangeLN nearly detects all change instances.

TABLE II
QUANTITATIVE RESULTS ON THE DSIFN. THE BEST RESULTS ARE MARKED IN BOLD, THE 2ND-BEST IS MARKED WITH UNDERLINE.

| Method | Backbone | Precision | Recall | F1 | IoU |
|---|---|---|---|---|---|
| FC-Siam-Diff | UNet | 60.09 | 61.39 | 60.74 | 43.61 |
| FC-Siam-Conc | UNet | 57.71 | 74.88 | 65.18 | 48.35 |
| STANet | ResNet18 | 84.61 | 88.62 | 86.57 | 76.32 |
| HANet | ResNet18 | 78.47 | 78.78 | 60.16 | 43.02 |
| IFN | VGG-16 | 90.00 | 95.40 | 92.62 | 76.26 |
| SNUNet | UNet++ | 87.22 | 83.90 | 85.53 | 74.71 |
| ChangeFormer | MiT-b1 | 93.13 | 93.59 | 93.36 | 87.55 |
| PCAM | ResNet18 | **93.98** | 92.34 | 93.15 | 87.18 |
| Changer | ResNet18 | 92.49 | 91.80 | 92.14 | 85.43 |
| ChangeLN | ResNet18 | 93.65 | 93.22 | 93.44 | 87.68 |
| ChangeLN* | MiT-b0 | 93.68 | **94.98** | **94.33** | **89.26** |

Quantitative results are presented in Table II. The DSIFN dataset, with its more complex scene compositions and diverse change instances beyond just buildings, better reflects the sensitivity of algorithms to pseudo-changes compared to the LEVIR-CD dataset. Compared to other methods, ChangeLN exhibits superior generalization performance, leading in combined F1 and IoU metrics. Specifically, ChangeLN_T, based on the Mit-b0 backbone, outperforms ChangeFormer by 0.97% in F1 score and 1.71% in IoU.

**Comparison Experiments on the SVCD:** The SVCD

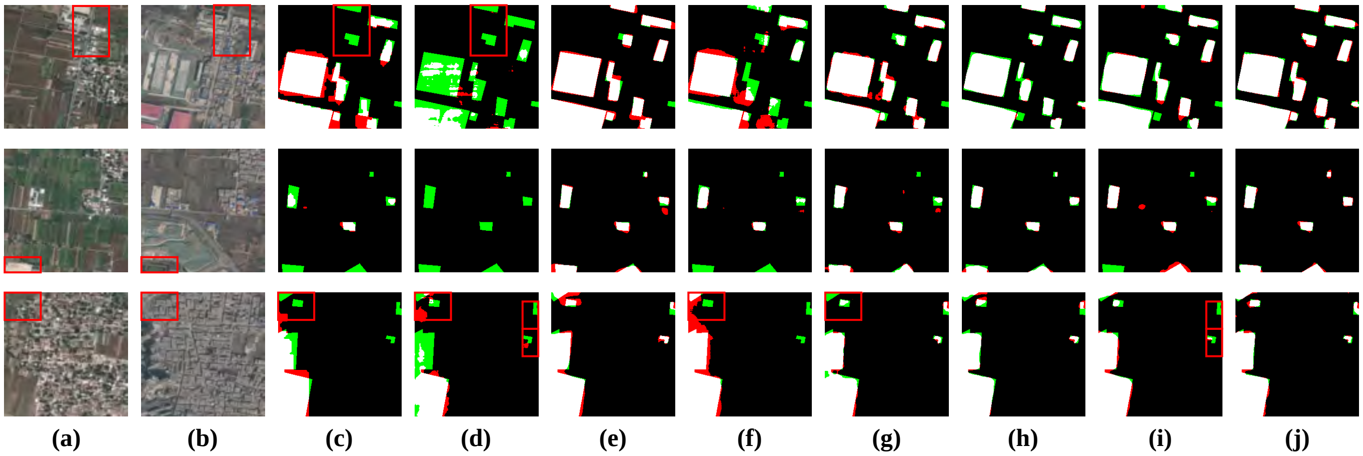

Fig. 6. The qualitative comparison of different methods on the DSIFN dataset. Please zoom-in for the best view. (a) Pre-temporal image. (b) Post-temporal image. (c) STANet. (d) HANet. (e) IFN. (f) SNUNet. (g) ChangeFormer. (h) PCAM. (i) Changer. (j) Ours. We highlight the TP areas in white, the FP areas in red, and the FN areas in green. The black color denotes the TN areas.

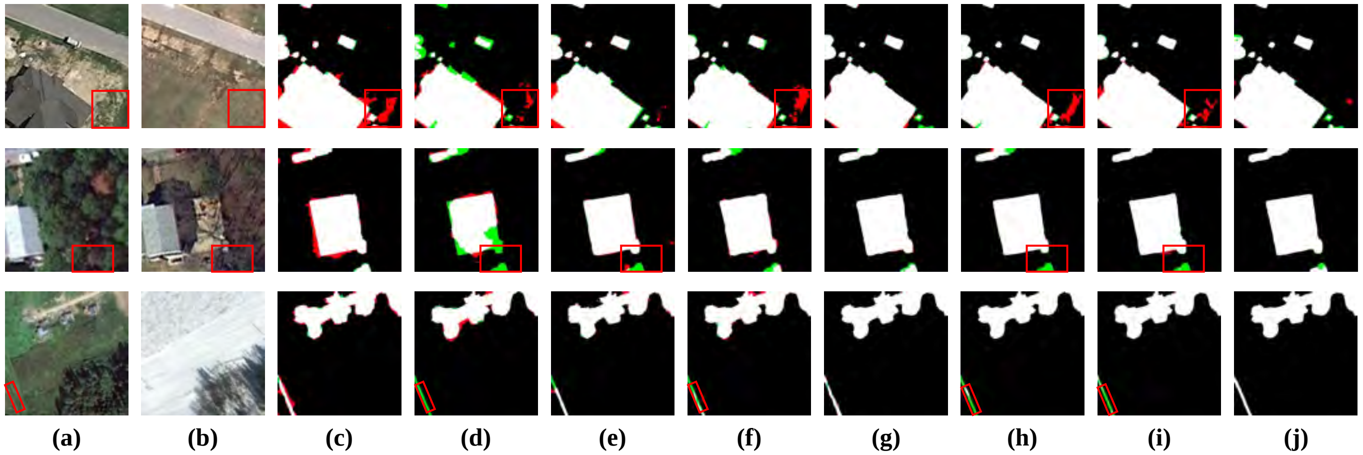

Fig. 7. The qualitative comparison of different methods on the SVCD dataset. Please zoom-in for the best view. (a) Pre-temporal image. (b) Post-temporal image. (c) STANet. (d) HANet. (e) IFN. (f) SNUNet. (g) ChangeFormer. (h) PCAM. (i) Changer. (j) Ours. We highlight the TP areas in white, the FP areas in red, and the FN areas in green. The black color denotes the TN areas.

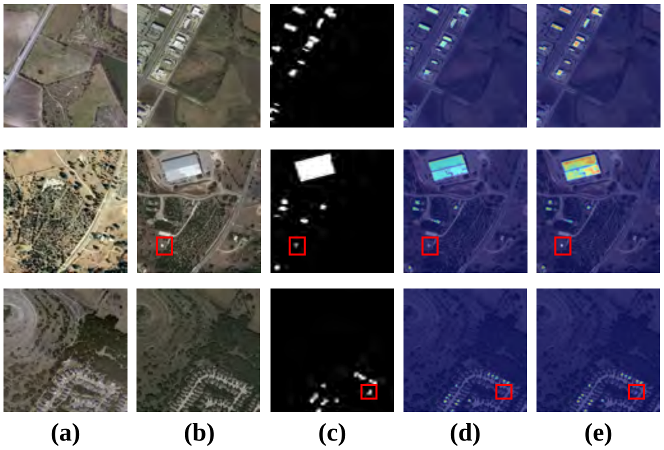

Fig. 8. Example of 3D-NCD module visualization by Gradient-weighted class activation maps (Grad-CAM). (a) Pre-temporal image. (b) Post-temporal image. (c) Ground truth. (d) Grad-CAM of the model without 3D-NCD module. (e) Grad-CAM of the model with 3D-NCD module.

dataset encompasses not only architectural transformations but also alterations in roads, vehicles, and vegetation. Characterized by lower image resolution, it presents more challenging and complex scenes, particularly due to typical seasonal differences, which compound interpretive difficulties. Fig. 7 illustrates the qualitative results of comparative methods on the SVCD dataset. In Row 1, the right side of the building, featuring bare land with spectral properties similar to rooftops, leads to some false detections in the contrast models. In Row 2, an change instance in the lower right-hand side, owing to poor imaging quality of remote sensing data, presents difficulties in feature extraction. Compared to its competitors, ChangeLN manages to more completely capture this change instance. Row 3 highlights an issue with grassy paths, which are not only inconspicuous but also entirely obscured in winter, resulting in missed detections in some models (e.g., PCAM and Changer) due to spectral similarities with the surroundings.

Table III presents the quantitative results for the SVCD

TABLE III
QUANTITATIVE RESULTS ON THE SVCD. THE BEST RESULTS ARE MARKED IN BOLD, THE 2ND-BEST IS MARKED WITH UNDERLINE.

| Method | Backbone | Precision | Recall | F1 | IoU |
|---|---|---|---|---|---|
| FC-Siam-Diff | UNet | 92.84 | 72.83 | 91.63 | 68.96 |
| FC-Siam-Conc | UNet | 90.46 | 72.06 | 80.22 | 66.97 |
| STANet | ResNet18 | 86.95 | 98.14 | 92.20 | 85.53 |
| HANet | ResNet18 | 92.18 | 86.20 | 89.09 | 80.33 |
| IFN | VGG-16 | 95.23 | 96.84 | 96.03 | 92.36 |
| SNUNet | UNet++ | 96.68 | 95.98 | 96.33 | 92.91 |
| ChangeFormer | MiT-b1 | 97.42 | 97.35 | 97.39 | 94.91 |
| PCAM | ResNet18 | 97.75 | 97.17 | 97.46 | 95.04 |
| Changer | ResNet18 | 97.71 | 97.18 | 97.44 | 95.02 |
| ChangeLN | ResNet18 | **98.06** | **97.78** | **97.92** | **95.92** |
| ChangeLN* | MiT-b0 | 97.70 | 97.06 | 97.38 | 94.89 |

TABLE IV
ABLATION EXPERIMENT ON THE LEVIR DATASET. THE BEST RESULTS ARE MARKED IN BOLD.

| 3D-NCD | DfR | Precision | Recall | F1 | IoU |
|---|---|---|---|---|---|
|  |  | 93.17 | 88.76 | 90.91 | 83.34 |
| ✔ |  | **93.24** | 90.41 | 91.81 | 84.85 |
|  | ✔ | 92.63 | 90.26 | 91.43 | 84.21 |
| ✔ | ✔ | 93.14 | **90.83** | **91.97** | **85.14** |

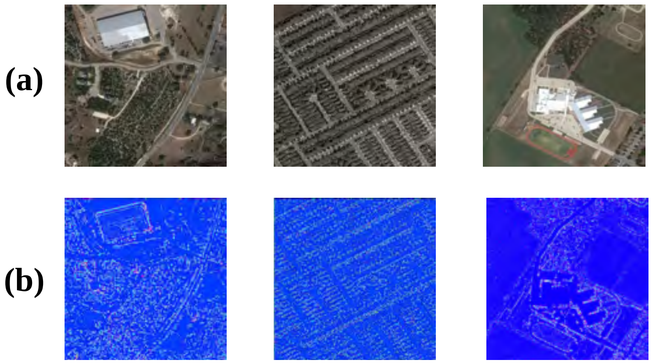

Fig. 9. The visualisation of Laplacian Details Detection. (a) Input image. (b) Detailed features detected by the Laplacian kernel.

TABLE V
QUANTITATIVE RESULTS OF TRANSFERRING THE 3D NEIGHBORHOOD CROSS DIFFERENCE STRATEGY TO OTHER MODELS.

| Method | Precision | Recall | F1 | IoU |
|---|---|---|---|---|
| FC-Siam-Conc | 57.71 | 74.88 | 65.18 | 48.35 |
| FC-Siam-3DNCD | 64.47 | 81.01 | 71.80 | 56.01 |
| STANet | 84.61 | 88.62 | 86.57 | 76.32 |
| STANet-3DNCD | 86.36 | 92.83 | 89.48 | 80.96 |
| ChangeFormer | 93.13 | 93.59 | 93.36 | 87.55 |
| ChangeFormer-3DNCD | 95.83 | 96.01 | 95.92 | 92.16 |
| Changer | 92.49 | 91.80 | 92.14 | 85.43 |
| Changer-3DNCD | 93.83 | 93.18 | 93.50 | 87.80 |

dataset. The subtle changes in roads, small-scale vehicle modifications, and spatial misalignments within the SVCD dataset impose heightened demands on the robustness of detection models. In terms of composite metrics F1 and IoU, ChangeLN maintains a lead of 0.48% and 0.9%, respectively, over Changer.

### D. Ablation Studies

To investigate the impact of the 3D Neighborhood Cross Difference strategy and the Detail-focused Refinement Decoder on the performance of ChangeLN, we conducted a comprehensive ablation study on the LEVIR dataset by comparing the effects of removing specific components. The quantitative results are provided in Table IV.

*1) Baseline:* The baseline network employs the same backbone as ChangeLN. For the modeling of differential features, the backbone features are concatenated and then processed using 2D convolution to extract differential features, in lieu of the 3D Neighborhood Cross Difference strategy. Furthermore, the Detail-focused Refinement Decoder is replaced by FFN decoder. To ensure a fair comparison, the baseline uses a cross-entropy loss function for supervision, and all other hyperparameters are maintained consistently across all comparative experiments. The baseline achieved an F1 score of 90.91% and an IoU of 83.34% on the LEVIR-CD dataset.

*2) Effects of 3D Neighborhood Cross Difference:* To mine the profound differences between bi-temporal images, we develop the 3D Neighborhood Cross Difference strategy to extract discriminative difference features. As presented in Table IV, the proposed difference module notably improves the performance of CD, as indicated in rows 2 and 4. Particularly, integrating the 3D Neighborhood Cross Difference strategy into the baseline results in a significant enhancement of the mIoU metric, elevating it from 83.34% to 84.85%.

To further substantiate the efficacy of the 3D Neighborhood Cross Differencing in the extraction of robust differential features, this study employs a visual comparison using Gradient-Weighted Class Activation Mapping (Grad-CAM) [48] with and without the incorporation of the 3D Neighborhood Cross Difference strategy. Models not utilizing this strategy rely on conventional convolution for the modeling of differential features. As illustrated in Fig. 8, compared to traditional convolution techniques, the proposed strategy demonstrates a more comprehensive extraction of change features. Moreover, it significantly enhances the areas of object changes, thereby underscoring its advantages in extracting distinctive differential features. In the final two rows, it is evident that for instances of subtle changes with less pronounced features, the 3D Neighborhood Cross Difference strategy robustly extracts their differential information, outperforming traditional convolutional approaches.

Table V presents the performance enhancement achieved by integrating the 3D Neighborhood Cross Difference Strategy into models with similar architectures, tested on the DSIFN dataset. The results indicate a substantial performance improvement across different models. Specifically, the strategy yielded performance increments in the comprehensive IoU metric by 7.66%, 4.64%, 4.61%, and 2.37% for FC-Siam-Conc, STANet, ChangeFormer, and Changer, respectively. Notably, the enhanced ChangeFormer significantly outperforms the proposed ChangeECD on this dataset. This superior performance can be attributed to the backbone of ChangeFormer, MiT-b1, which due to its advantageous parameter quantity, exhibits exceptional capability in fitting and capturing land cover changes.

*3) Effects of Detail-focused Refinement Decode:* To enhance the detail representation of change instances, we have developed a Detail-focused Refinement Decoder based on the Laplacian operator. As indicated in Table IV, various ablation tests involving different combinations of components show that the DfR module contributes an improvement ranging from 0.19 to 1.22 percentage points. Fig. 9 visualizes the gradient information extracted by the Laplacian Detail Detection Unit. This information is integrated through an attention mechanism with differential features derived from the 3D Neighborhood Cross Difference Module, thereby enriching the detail representation of the change maps.

## V. Conclusion

In this paper, we introduce a novel difference feature extraction strategy, referred to as 3D Neighborhood Cross Differencing, for exploring deep-level disparities in spatiotemporal instance features. Furthermore, the Detail-focused Refinement Decoder is constructed to enhance the fine-grained representation of change maps. Experimental results demonstrate that the proposed ChangeLN outperforms SOTA methods in both quantitative metrics and qualitative results across three public datasets. Simultaneously, the 3D Neighborhood Cross Differencing is transplanted into other similar architectures, resulting in significant performance improvements. To further enhance the robustness of differential feature modeling, the future research will focus on how to map bi-temporal features into a unified feature space to mitigate the challenges associated with differential modeling.

## References

[1] Z. Li, S. Cao, J. Deng, F. Wu, R. Wang, J. Luo, and Z. Peng, "Stade-cdnet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.

[2] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, 2005.

[3] D. Lu, P. Mausel, E. BrondĂzio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.

[4] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 228–239, 2022.

[5] A. A. Abuelgasim, W. Ross, S. Gopal, and C. Woodcock, "Change detection using adaptive fuzzy neural networks: Environmental damage assessment after the gulf war," *Remote Sens. Environ.*, vol. 70, no. 2, pp. 208–223, 1999.

[6] G. Satalino, F. Mattia, A. Balenzano, F. P. Lovergine, M. Rinaldi, A. P. De Santis, S. Ruggieri, D. A. Nafría García, V. P. Gómez, E. Ceschia, M. Planells, T. L. Toan, A. Ruiz, and J. Moreno, "Sentinel-1 & sentinel-2 data for soil tillage change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2018, pp. 6627–6630.

[7] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, 2016.

[8] R. A. Weismiller, S. J. Kristof, D. K. Scholz, P. E. Anuta, and S. Momin, "Change detection in coastal zone environments," *Photogramm. Eng. Remote Sensing*, vol. 43, 1977.

[9] P. Coppin, I. Jonckheere, K. Nackaerts, and et al., "Digital change detection methods in ecosystem monitoring: a review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[10] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using tri-temporal logic-verified change vector analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 278–293, 03 2020.

[11] 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.

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

[13] L. Bai, W. Huang, X. Zhang, S. Du, G. Cong, H. Wang, and B. Liu, "Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois," *ISPRS J. Photogramm. Remote Sens.*, vol. 201, pp. 193–208, 2023.

[14] Q. Zhu, Y. Zhang, L. Wang, Y. Zhong, Q. Guan, X. Lu, L. Zhang, and D. Li, "A global context-aware and batch-independent network for road extraction from vhr satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 353–365, 2021.

[15] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[16] Y. Yuan, Z. Li, and D. Ma, "Feature-aligned single-stage rotation object detection with continuous boundary," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[17] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[18] 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.

[19] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, 2020.

[20] M. Lebedev, Y. Vizilter, O. Vygolov, V. Knyaz, and A. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-2, pp. 565–571, 05 2018.

[21] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, 2019.

[22] S. Sun, L. Mu, L. Wang, and P. Liu, "L-unet: An lstm network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

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

[24] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5294–5304.

[25] X. Ning, H. Zhang, R. Zhang, and X. Huang, "Multi-stage progressive change detection on high resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 207, pp. 231–244, 2024.

[26] Y. Cao, X. Huang, and Q. Weng, "A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas," *Remote Sensing of Environment*, vol. 297, p. 113779, 2023.

[27] W. Jing, Y. Yuan, and Q. Wang, "Dual-field-of-view context aggregation and boundary perception for airport runway extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[28] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 31, pp. 7252–7263, 2022.

[29] M. Lin, G. Yang, and H. Zhang, "Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images," *IEEE Trans. Image Process.*, vol. 32, pp. 57–71, 2023.

[30] X. Zhang, M. Tian, Y. Xing, Y. Yue, Y. Li, H. Yin, R. Xia, J. Jin, and Y. Zhang, "Adhr-cdnet: Attentive differential high-resolution change detection network for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[31] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, "Difference enhancement and spatial–spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10 012–10 022.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[35] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1489–1500, 2023.

[36] F. Zhu, J. Cui, and K. Dou, "Spatio-temporal hierarchical feature transformer for uav object tracking," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 442–452, 2023.

[37] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

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

[39] R. Huang, R. Wang, Q. Guo, Y. Zhang, and W. Fan, "Idet: Iterative difference-enhanced transformers for high-quality change detection," 2022.

[40] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[41] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, "Multiscale diff-changed feature fusion network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.

[42] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and X. Gao, "Dcenet: Diff-feature contrast enhancement network for semi-supervised hyperspectral change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[43] Y. Zhao, P. Chen, Z. Chen, Y. Bai, Z. Zhao, and X. Yang, "A triple-stream network with cross-stage feature fusion for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

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

[45] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[46] S. Wang, Y. Li, M. Xie, M. Chi, Y. Wang, C. Wang, and W. Zhu, "Align, perturb and decouple: Toward better leverage of difference information for rsi change detection," 08 2023, pp. 1497–1505.

[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, pp. 1–11, 2023.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

**Wei Jing** received the B.M. degree in e-commerce and the M.S. degree in computer software and theory from Shandong University of Science and Technology, Qingdao, China, in 2019 and 2022 respectively.

He is currently working toward the Ph.D. degree in the National Elite Institute of Engineering and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing image processing and deep learning.

**Kaichen Chi** received the B.E. degree in electronic and information engineering and the M.E. degree in communication and information system from Liaoning Technical University, Huludao, China, in 2019 and 2022 respectively. He is currently working toward the Ph.D. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include image processing and deep learning.

**Qiang Li** is currently with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

**Qi Wang** (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the Uni- versity of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.