# TriCDNet: A Multi-scale Tri-stream Interaction Network for Building Change Detection

Renjie Zhou

School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China – zhourenjie12@163.com

**Keywords:** Building Change Detection, Remote Sensing, Feature Interaction, Transformer, Multi-scale Fusion.

**Abstract**

Building change detection (BCD) from multi-temporal remote sensing imagery plays a vital role in urban monitoring and land management. However, existing deep learning-based methods still suffer from insufficient semantic differentiation, weak multi-scale consistency, and limited global context modeling. To address these issues, we propose TriCDNet, a multi-scale tri-stream interaction network for accurate and robust building change detection. The network integrates three complementary feature streams—bi-temporal features and their normalized difference map—and performs stage-wise feature interaction through a multi-layer Difference-guided Cross-temporal Interaction Module (DCIM). A top-down Feature Pyramid Network (FPN) is employed to aggregate multi-scale information, while a lightweight Transformer-based decoder captures long-range spatial–temporal dependencies for global reasoning. Experiments on three public datasets (LEVIR-CD, WHU-CD, and SYSU-CD) demonstrate that TriCDNet achieves superior accuracy and structural consistency, with IoU of 85.2%, 86.91%, and 71.54%, respectively. The results confirm that each component contributes positively to performance, and the proposed tri-stream framework effectively balances local detail preservation and global semantic coherence, showing strong generalization capability in complex urban scenes.

## 1. Introduction

With the rapid development of remote sensing technologies, multi-temporal high-resolution imagery has become increasingly accessible, providing strong data support for dynamic surface monitoring and change analysis. Change detection (CD) aims to identify differences within the same area across images captured at different times and has been widely applied in disaster assessment (Zheng et al., 2021), environmental monitoring (Wu et al., 2020), and land management (Lv et al., 2021). However, achieving efficient and accurate CD in complex scenarios remains a major challenge.

Early change detection methods primarily relied on handcrafted feature differencing and threshold-based techniques (He et al., 2014) or principal component analysis (Munyati, 2004), which achieved limited performance in complex scenes. Although these approaches are simple to implement and can yield reasonable results in specific cases, they often fail to handle complex backgrounds and fine-grained targets. With the rise of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm in change detection, owing to their strong feature extraction capability. Built upon encoder–decoder architectures, CNN-based methods effectively learn hierarchical representations from bi-temporal images. In particular, Siamese networks have been widely adopted as the prevailing framework for modeling bi-temporal relationships. Within this paradigm, various backbone architectures have been explored: early studies commonly used classical convolutional networks such as VGG (Simonyan and Zisserman, 2014) and ResNet (He et al.,2016); encoder–decoder structures like U-Net (Ronneberger et al., 2015) and its variants have been widely employed for dense prediction tasks; Transformer-based backbones such as the Vision Transformer (ViT) (Dosovitskiy, 2020)and Swin Transformer(Liu et al., 2021) have been introduced to enhance global modeling capacity; and lightweight models such as EfficientNet (Tan and Le, 2019) have attracted attention for their balance between accuracy and efficiency, making them suitable for large-scale applications. Building upon these architectural evolutions, researchers have developed a series of Siamese-based models that further refine bi-temporal representation and fusion strategies.Daudt et al. (2018) were the first to apply Siamese networks to remote sensing change detection, introducing two fully convolutional models—FC-Siam-conc and FC-Siam-diff, which perform change modeling via feature concatenation and differencing, respectively. These models marked the early use of CNN architectures in this field. Building on this foundation, subsequent studies explored more advanced feature fusion strategies across multiple branches. Fang et al. (2023) enhanced bi-temporal feature alignment and fusion by incorporating a feature interaction layer and a Flow Dual Alignment Fusion (FDAF) module within the Meta Changer framework, demonstrating the importance of feature interaction in improving detection accuracy. Han et al. (2023) proposed CG-Net, which derives change maps from semantically rich deep features and uses them as priors to guide multi-scale fusion. Nevertheless, due to the inherently limited receptive field of convolutional networks, these methods struggle to capture long-range contextual dependencies. To mitigate this, Chen et al. (2022) introduced Transformer modules into the change detection pipeline, boosting global modeling capacity and enhancing the representation of semantic changes. Bandara and Patel (2022) proposed a Transformer-based Siamese change detection network (ChangeFormer), which combines a Transformer encoder with multilayer perceptrons in a Siamese architecture to effectively capture multi-scale long-range details for improved accuracy. Although prior studies have introduced feature interaction and global modeling mechanisms, they still fall short in semantic differencing, multi-scale structural consistency, and contextual information integration. To address the above challenges, this paper proposes a multi-scale tri-stream interaction framework for building change detection. A normalized difference stream highlights semantic variations between bi-temporal features, and the three streams are fused through a Difference-guided Cross-temporal Interaction Module (DCIM) that performs bidirectional interaction and gated fusion

across multiple scales. Aggregated features are refined by a Feature Pyramid Network (FPN) for semantic consistency, and a Bi-temporal Transformer (BIT) models long-range spatial–temporal dependencies. Experiments on multiple public datasets demonstrate the effectiveness and robustness of the proposed framework. The main contributions of this work are summarized as follows:

(1) A tri-stream feature interaction mechanism based on the proposed DCIM for fine-grained temporal correspondence.

(2) Multi-scale feature aggregation via FPN to maintain semantic and structural consistency.

(3) Global modeling with BIT to capture long-range dependencies and enhance semantic representations.

## 2. Methodology

### 2.1 Overall Architecture

In this paper, we have designed a novel change detection network featuring a tri-stream interaction architecture, as illustrated in Figure 1. The model comprises four main components: a shared-weight feature extraction backbone based on EfficientNet-B5, a multi-level Difference-guided Cross-temporal Interaction Module (DCIM), a Feature Pyramid Network (FPN), and a Transformer-based decoder (BIT). The backbone is responsible for extracting hierarchical representations from bi-temporal input images. To enhance the model's sensitivity to changed regions, a third feature stream is introduced, which is generated through normalized differencing between the bi-temporal feature maps. At each feature level, the three streams are fed into the corresponding DCIM module to strengthen cross-temporal feature interaction. The interacted features are then aggregated by the FPN to achieve multi-scale fusion. The fused multi-scale features are transformed into tokens and passed into a Transformer encoder–decoder structure to model long-range spatial dependencies. Finally, the model computes the absolute difference between the decoded features and fuses it with the difference-guided branch features, followed by a convolutional classifier to generate the final change mask.

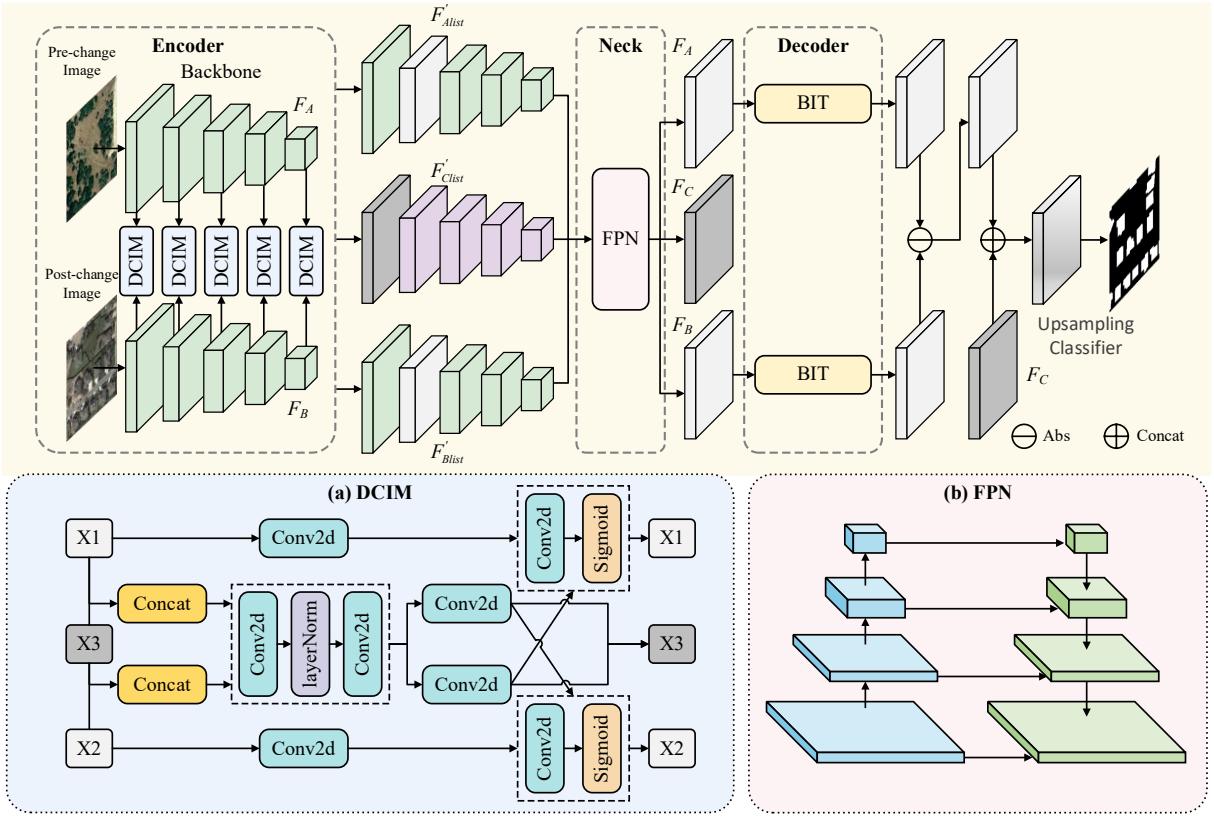


Figure 1. The architecture of the proposed TriCDNet.

### 2.2 EfficientNet-B5

In this study, we adopt EfficientNet-B5 as the feature extraction backbone of the change detection network. The EfficientNet series is designed based on the compound scaling strategy, which jointly scales network depth, width, and input resolution to enhance feature representation while maintaining a lightweight architecture. Unlike conventional convolutional neural networks that expand along a single dimension, the compound scaling mechanism enables EfficientNet to achieve superior performance with fewer parameters.

For remote sensing change detection tasks that focus on local structural differences, EfficientNet-B5 provides a favourable balance between high-resolution spatial perception and strong global modeling capability, effectively capturing fine-grained change features in the imagery. Therefore, EfficientNet-B5 is employed as the feature extractor to provide high-quality semantic representations for subsequent multi-scale feature interaction.

### 2.3 Tri-stream Feature Interaction

To effectively capture change information between bi-temporal remote sensing images and enhance the model's sensitivity to subtle local differences, we design a multi-layer Difference-guided Cross-temporal Interaction Module (DCIM) to enable multi-scale tri-stream feature interaction. The module integrates the first-temporal features, second-temporal features, and their similarity-guided difference features, reinforcing

cross-temporal collaborative representation while preserving the discriminative capacity of each feature stream.

**2.3.1 Input Design**: At each stage, the backbone network outputs bi-temporal feature maps $F_i^{t1}$ and $F_i^{t2}$ . To highlight potential change regions, a difference prior feature $F_i^{\Delta}$ is introduced and computed as follows:

$$F_i^{\Delta} = 1 - \frac{(F_i^{t1} - F_i^{t2})^2}{1 + (F_i^{t1} - F_i^{t2})^2}. \quad (1)$$

This normalized differencing suppresses noise and enhances sensitivity to subtle structural variations. Consequently, each DCIM receives three input streams: $\{F_i^{t1}, F_i^{t2}, F_i^{\Delta}\}$ .

**2.3.2 Cross-temporal Interaction Mechanism**: To effectively exploit the complementary information between bi-temporal features and improve the model's sensitivity to subtle structural variations, we design a Difference-guided Cross-temporal Interaction Module (DCIM) inspired by the Bilateral Information Exchange (BIE) mechanism (Huang et al., 2023). While the original BIE was developed for event stream super-resolution and implemented an attention-based bilateral information propagation, our DCIM extends this concept to a convolution-driven tri-stream framework for remote sensing change detection. It introduces an additional difference-guided stream to explicitly encode potential change priors and strengthen cross-temporal consistency.First, the temporal features $F_i^{t1}$ and $F_i^{t2}$ are refined through residual convolutional blocks to strengthen semantic representation and spatial consistency. Then, the difference prior $F_i^{\Delta}$ is concatenated with each temporal feature and compressed via lightweight 1×1 convolutions and normalization, forming a difference-guided joint feature space for subsequent interaction.Following the conceptual design of BIE, we compute cross-correlations between the two temporal branches through batch matrix multiplication, which acts as a lightweight attention-style operator that enables bidirectional semantic information exchange under the guidance of the difference stream.

To adaptively integrate the enhanced and original features, a gated fusion mechanism is employed. Each temporal feature passes through a 1×1 convolution and a sigmoid activation to generate a gating map, which dynamically balances the contributions of change-sensitive and stable components:

$$G_i^{t1} = \sigma(\mathrm{Conv}_{1\times1}(F_i^{t1*} + F_i^{t1})), \quad (2)$$

$$F_i^{t1'} = G_i^{t1} \odot F_i^{t1*} + (1 - G_i^{t1}) \odot F_i^{t1}. \quad (3)$$

This gating mechanism adaptively balances change-sensitive and stable components, enabling the model to emphasize meaningful structural variations while suppressing background noise. Finally, contextual information is aggregated by concatenating the temporal representations and integrating them with the difference feature through a residual connection, yielding a comprehensive difference-guided representation $F_i^{\Delta'}$ . Through this progressive attention and fusion process, DCIM effectively strengthens cross-temporal dependency modeling and enhances the network's capability to capture localized building changes.

**2.3.3 Learnable Reweighting and Multi-level Embedding**: To further improve adaptability, learnable scalar parameters $\alpha_i$ and $\beta_i$ (initialized to 0.5) are introduced to balance interaction and contextual information during training. The reweighted features are computed as follows:

$$\tilde{F}_i^{t1} = \alpha_i F_i^{t1'} + (1 - \alpha_i) F_i^{\Delta'}, \tilde{F}_i^{t2} = \beta_i F_i^{t2'} + (1 - \beta_i) F_i^{\Delta'}. \quad (4)$$

This reweighting mechanism enables the network to dynamically control the trade-off between cross-temporal interaction and contextual consistency. As a result, it maintains semantic alignment while enhancing change sensitivity. DCIM modules are embedded at five stage of the backbone $(i = 0,1,2,3,4)$ , forming a progressive, multi-level interaction pathway. Each level outputs three updated features $\{\tilde{F}_i^{t1}, \tilde{F}_i^{t2}, \tilde{F}_i^{\Delta}\}$ , which are subsequently aggregated by the Feature Pyramid Network (FPN) for cross-scale fusion. The fused multi-scale representations are further refined through a Transformer-based decoder to model global dependencies and generate the final pixel-level change map.

### 2.4 Global Context Modeling and Prediction Head

While the DCIM modules focus on local feature interaction and difference enhancement, the subsequent components aim to capture global dependencies and generate the final change map. This stage consists of three submodules: a multi-scale fusion network (FPN), a Transformer-based interaction decoder (BIT), and a change prediction head. Together, they form the high-level semantic reasoning and output stage of the proposed model.

**2.4.1 Feature Pyramid Network (FPN)**: To effectively aggregate the multi-scale features generated by the DCIM modules, we employ a modified Feature Pyramid Network (FPN) as the neck structure. Each DCIM block outputs three feature streams—pre-change temporal, post-change temporal, and difference-guided—denoted as $\{\tilde{F}_i^{t1}, \tilde{F}_i^{t2}, \tilde{F}_i^{\Delta}\}$ . These features are extracted from five hierarchical stages of the backbone corresponding to spatial resolutions of 1/2, 1/4, 1/8, 1/16, and 1/32 of the 256×256 input image. The feature maps have sizes of 128×128, 64×64, 32×32, 16×16, and 8×8, with channel dimensions of [24, 40, 64, 176, 512].

All features are first projected to 128 channels using lateral 1×1 convolutions to ensure consistent dimensionality across scales before multi-level fusion. The FPN constructs a top-down pathway with lateral skip connections, progressively upsampling and merging high-level semantic representations with low-level spatial details. After fusion, three aggregated feature maps are obtained, corresponding to the two temporal streams and the difference-guided stream. Specifically, the processed features have spatial dimensions as:

$$F_{\mathrm{FPN}}^{t1} \in \mathbb{R}^{128\times64\times64}, F_{\mathrm{FPN}}^{t2} \in \mathbb{R}^{128\times64\times64}, F_{\mathrm{FPN}}^{\Delta} \in \mathbb{R}^{128\times128\times128}. \quad (5)$$

The two temporal features retain semantic consistency from the DCIM, while the difference-guided stream preserves local change sensitivity. This multi-stream fusion strategy effectively bridges fine-grained spatial details and high-level semantics, providing comprehensive multi-scale representations for subsequent global reasoning in the Transformer decoder.

**2.4.2 Bitemporal Image Transformer (BIT)**: After multi-scale fusion, the semantically enhanced features are fed into the BIT [Chen *et al.*, 2022], which serves as a high-level decoder for global context reasoning and cross-temporal interaction. Unlike convolutional operations that are inherently local, BIT utilizes Transformer-based self- and cross-attention to capture long-range dependencies between the two temporal streams, allowing bidirectional semantic communication across time. In our implementation, we deploy BIT as an external decoder operating on the FPN outputs . These features are first tokenized through an attention-based pooling mechanism to obtain compact semantic tokens that encode essential contextual information. The tokens from both temporal branches are concatenated and passed into the Transformer encoder to model global correlations, which are then decoded back into the spatial domain to generate semantically enhanced feature representations.

**2.4.3 Change Prediction Head**: After Transformer decoding, the bi-temporal features are differenced and fused with the difference-guided stream to generate the final change prediction. Specifically, the outputs of BIT are first differenced to highlight potential change regions, while the difference-guided feature from the FPN is downsampled to obtain a compact local representation. These two features are then concatenated along the channel dimension, combining global semantic context from the Transformer decoder with locally guided difference cues. The fused representation is upsampled to recover spatial resolution and further refined by a lightweight doubleconv block, followed by a 1×1 convolution and sigmoid activation to produce the binary change map. This design effectively integrates global reasoning and local detail enhancement, improving both the accuracy and spatial precision of the detected changes.

## 3. Experiment

### 3.1 Dataset

We evaluate our method on three publicly available change detection datasets: LEVIR-CD, WHU-CD, and SYSU-CD.

LEVIR-CD (Chen and Shi, 2020a) is a large-scale dataset specifically curated for building change detection tasks. It comprises 637 pairs of high-resolution (0.5 meters/pixel) satellite images obtained from Google Earth, each with spatial dimensions of 1024×1024 pixels. Every image pair is annotated to accurately distinguish between changed and unchanged building regions, encompassing more than 31,000 individual change instances. Following the standard data split protocol, we use 445, 64, and 128 image pairs for training, validation, and testing, respectively. To facilitate training and reduce computational overhead, all images are cropped into non-overlapping 256×256 patches without any padding. This preprocessing step results in 7,120 training, 1,024 validation, and 2,048 testing patches.

WHU-CD (Ji et al., 2018) is a high-resolution building change detection dataset constructed from a pair of ultra-large aerial images captured in Christchurch, New Zealand, before and after a major earthquake in February 2011. These images, with a spatial resolution of 0.075 meters/pixel, span an area exceeding 30,000 × 15,000 pixels, and collectively contain approximately 21.4 million changed pixels and 481.9 million unchanged pixels. We follow the official data split provided on the project website, using 1,260 image pairs for training and 690 for testing. The original images are further divided into non-overlapping 256×256 patches for model input. A validation set is constructed by randomly selecting 10% of the training data, yielding 4,536 training, 504 validation, and 2,760 testing patches.

SYSU-CD (Shi et al., 2021) is a large-scale, category-agnostic change detection dataset that includes 20,000 pairs of aerial image patches with a spatial resolution of 0.5 meters/pixel and patch size of 256×256 pixels. Collected in Hong Kong between 2007 and 2014, the dataset captures a diverse range of change scenarios, such as urban development, suburban sprawl, groundwork operations, vegetation dynamics, road extensions, and coastal construction. The dataset is split according to a fixed 6:2:2 ratio, resulting in 12,000 training, 4,000 validation, and 4,000 testing pairs. Its diversity and large scale make it an effective benchmark for evaluating both general-purpose and structure-sensitive CD methods.

### 3.2 Implementation Details

All models are implemented using the PyTorch framework based on the MM Segmentation library. Experiments are conducted on an Ubuntu system with a single NVIDIA RTX 4090 GPU. The AdamW optimizer is adopted with an initial learning rate of 0.0003 and weight decay of 0.01. The learning rate follows a linear warm-up for 1,500 iterations and polynomial decay thereafter until 30,000 iterations. The batch size is 12, and data augmentation includes random rotation, flipping, and photometric distortion. Model selection is based on the best mIoU on the validation set.

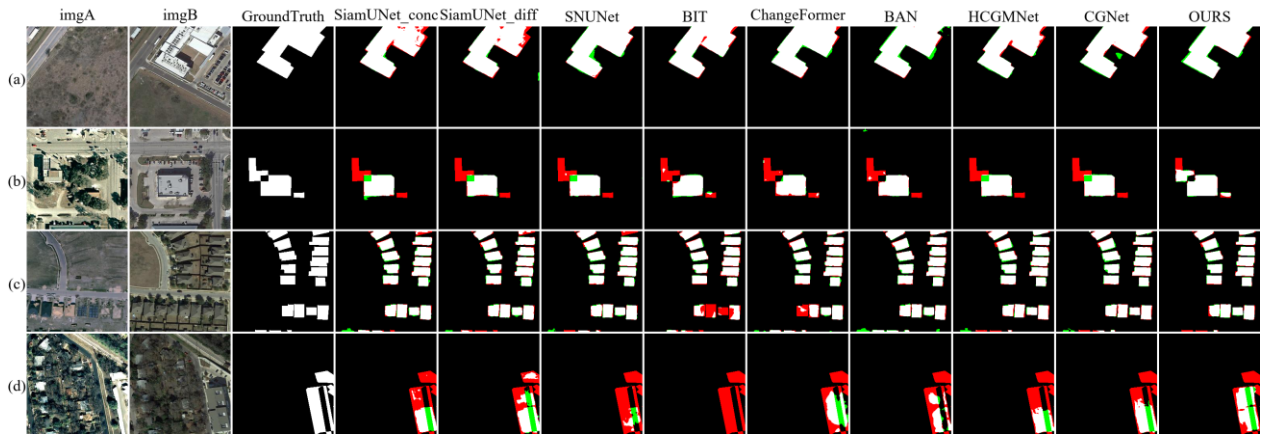


Figure 2. Visual comparison of change detection results on the LEVIR-CD. White, black, green, and red pixels denote true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), respectively.

### 3.3 Metrics

We evaluate the model performance using four standard metrics: Intersection over Union (IoU), F1-score, Precision, and Recall. All metrics are computed at the pixel level based on binary change masks, where the scores correspond to the change (foreground) class rather than the mean across classes. Their calculation formula is as follows:

$$IoU = \frac{TP}{TP + FN + FP}, \tag{6}$$

$$F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}, \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{8}$$

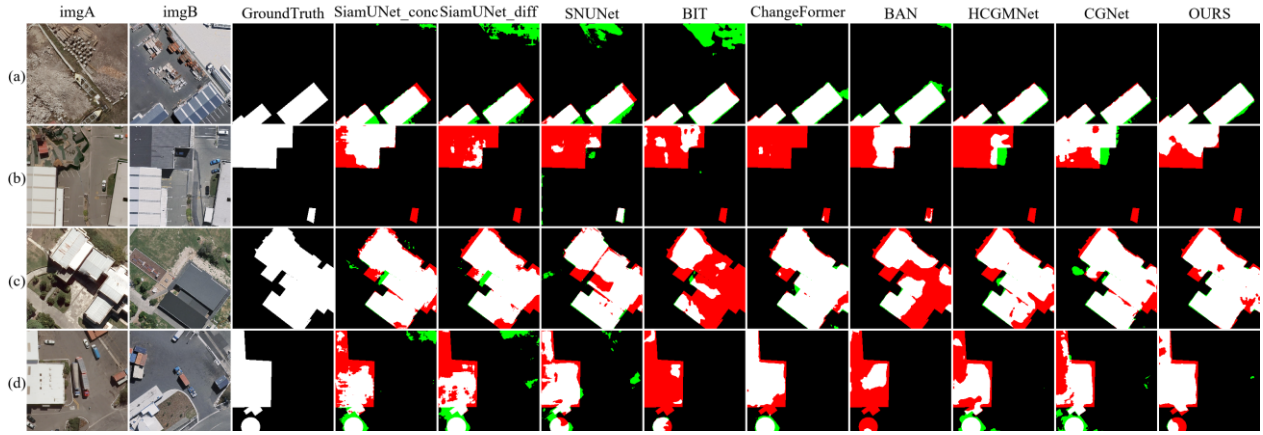


Figure 3. Visualization results of different methods on the WHU-CD dataset.

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{9}$$

### 3.4 Comparative Experiments

To comprehensively demonstrate the effectiveness of the proposed Tri-CD framework, we conducted both qualitative and quantitative comparisons with several state-of-the-art change detection (CD) methods. Specifically, the compared methods include Fully Convolutional Siamese-Concatenation (FC-Siam-conc) (Daudt et al., 2018), Fully Convolutional Siamese-Difference (FC-Siam-diff) (Daudt et al., 2018), Siamese NestedUNet (SNUNet) (Chen et al., 2022a), Bitemporal Image Transformer (BIT) (Chen et al., 2021), ChangeFormer (Bandara and Patel, 2022, Hierarchical Cross-Guided Multi-scale Network (HCGMNet) (Han et al., 2023b) ), Change Guiding Network (CGNet) (Han et al., 2023a), and Bi-Temporal Adapter Network (BAN) (Li et al., 2024b).

**Table 1**. Quantitative comparison of different methods on the LEVIR-CD dataset.

| Model | LEVIR-CD | | | |
|---|---|---|---|---|
| | IoU | Recall | F1 | Precision |
| FC-Siam-conc | 82.13 | 88.67 | 90.19 | 91.76 |
| FC-Siam-diff | 82.16 | 88.56 | 90.21 | 91.92 |
| SNUNet | 82.26 | 88.08 | 90.27 | 92.57 |
| BIT | 83.47 | 88.83 | 90.99 | 93.27 |
| ChangeFormer | 83.98 | 89.92 | 91.29 | 92.71 |
| BAN | 84.19 | 89.94 | 91.41 | 92.93 |
| HCGMNet | 84.79 | 90.61 | 91.77 | 92.96 |
| CGNet | 85.21 | 90.9 | 92.01 | 93.15 |
| **OURS** | 85.2 | 91.13 | 92.01 | 92.91 |

on three publicly available datasets: LEVIR-CD, WHU-CD, and SYSU-CD. Since this is a binary change detection task, the Intersection over Union (IoU) of the change (foreground) class was adopted as the primary evaluation metric.

On the LEVIR-CD dataset (Table 1), our model achieved outstanding performance with an IoU of 85.20% and an F1-score of 92.01%, performing on par with and slightly surpassing the best-performing CGNet. It also significantly outperformed classical models such as the FC-Siam series and BIT, demonstrating its strong capability in fine-grained building change recognition.

**Table 2**. Quantitative comparison of different methods on the WHU-CD dataset.

| Model | WHU-CD | | | |
|---|---|---|---|---|
| | IoU | Recall | F1 | Precision |
| FC-Siam-conc | 82.83 | 89.14 | 90.61 | 92.13 |
| FC-Siam-diff | 82.86 | 88.74 | 90.62 | 92.6 |
| SNUNet | 81.93 | 89.48 | 90.07 | 90.66 |
| BIT | 83.97 | 88.08 | 91.29 | 94.74 |
| ChangeFormer | 84.49 | 89.28 | 91.6 | 94.04 |
| BAN | 85.22 | 88.72 | 92.02 | 95.57 |
| HCGMNet | 85.33 | 90.31 | 92.08 | 93.93 |
| CGNet | 86.21 | 90.79 | 92.59 | 94.47 |
| **OURS** | 86.91 | 90.39 | 93 | 95.77 |

On the WHU-CD dataset (Table 2), our model further exhibited excellent cross-domain generalization, achieving an IoU of 86.91% and an F1-score of 93.00%, outperforming recent methods such as CGNet and BAN across all evaluation metrics. In particular, it achieved a Precision of 95.77%, indicating a clear advantage in reducing false positives.

On the SYSU-CD dataset (Table 3), TriCDNet continued to maintain superior performance, reaching an IoU of 71.54% and an F1-score of 83.41%, which represents an improvement of approximately 2.4 percentage points in IoU over the second-best method, BAN. These results demonstrate that the proposed network maintains strong robustness in complex multi-class change detection scenarios and performs particularly well in urban environments and fine-grained target recognition.

To further validate the model's performance across different scenes, we conducted visual comparisons on representative samples from the three datasets (Figs. 1–3). In the visualizations, white pixels denote true positives (TP), black pixels true negatives (TN), green pixels false positives (FP), and

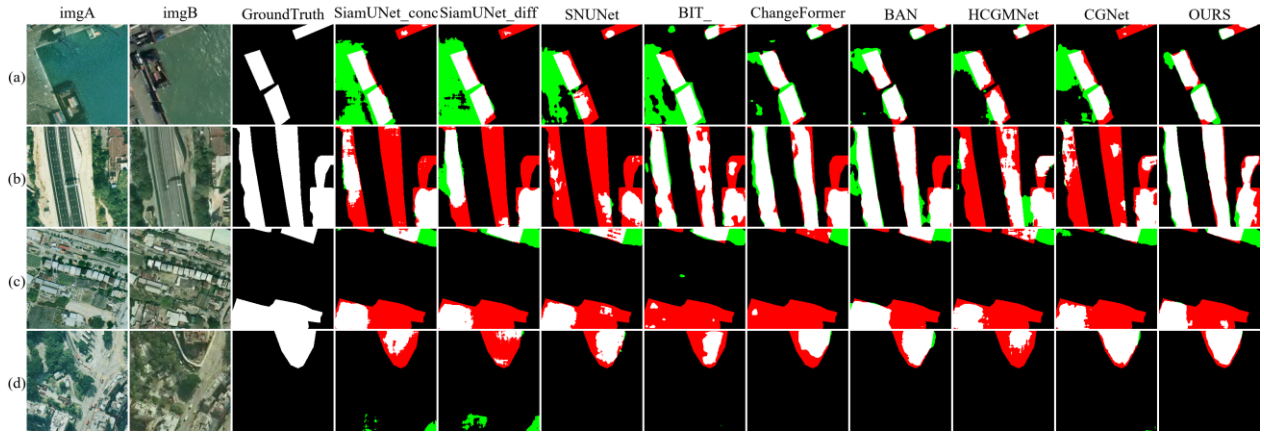


Figure 4. Visualization results of different methods on the SYSU-CD dataset.

red pixels false negatives (FN). As shown in the results, our method better preserves structural integrity along object boundaries and effectively suppresses both false alarms and missed detections.

**Table 3**. Quantitative comparison of different methods on the SYSU-CD dataset.

| Model | SYSU-CD | | | |
|---|---|---|---|---|
| | IoU | Recall | F1 | Precision |
| FC-Siam-conc | 65.39 | 75.39 | 79.07 | 83.14 |
| FC-Siam-diff | 65.74 | 75.2 | 79.33 | 83.94 |
| SNUNet | 65.49 | 75.11 | 79.14 | 83.63 |
| BIT | 65.35 | 74.52 | 79.04 | 84.14 |
| ChangeFormer | 67.87 | 77.4 | 80.86 | 84.65 |
| BAN | 68.08 | 79.05 | 81.01 | 83.06 |
| HCGMNet | 66.33 | 74.15 | 79.76 | 86.28 |
| CGNet | 66.55 | 74.37 | 79.92 | 86.37 |
| **OURS** | 71.54 | 80 | 83.41 | 87.13 |

scores overall. The complete TriCDNet model achieves IoU = 85.2% and F1 = 92.0% on LEVIR-CD, IoU = 86.9% and F1 = 93.0% on WHU-CD, and IoU = 71.5% and F1 = 83.4% on SYSU-CD, outperforming all intermediate variants.

These quantitative results demonstrate that each module contributes positively to the final performance, and the combined architecture effectively balances local detail preservation and global semantic consistency.

Although TriCD achieves clear improvements over baseline variants, slight performance variations remain across datasets, likely due to the reliance on backbone capacity and the sensitivity of Transformer modules to scale differences. Future work will focus on developing more efficient interaction mechanisms and adaptive normalization strategies to further enhance generalization.

### 3.5 Ablation Studies

To verify the effectiveness of each component in the proposed TriCDNet framework, we conducted a series of ablation experiments on the LEVIR-CD, WHU-CD, and SYSU-CD datasets. The baseline model adopts only the EfficientNet-B5 backbone and BIT decoder, while the other modules are incrementally added for comparison. Specifically, B5 denotes the backbone used for feature extraction; DCIM refers to the Difference-guided Cross-temporal Interaction Module; FPN represents the Feature Pyramid Network for multi-scale feature aggregation; and BIT indicates the Bitemporal Image Transformer for global reasoning.

As presented in Table 4, adding DCIM leads to consistent performance improvements across all datasets, confirming its effectiveness in enhancing cross-temporal interaction and reducing false detections. The introduction of FPN further increases both IoU and F1 scores, highlighting the importance of multi-scale feature fusion for detailed change representation. When the BIT decoder is integrated, the model benefits from global context reasoning, yielding the highest

**Table 4**. Ablation study of the proposed TriCDNet framework on three datasets. The best results are highlighted in bold. A represents the EfficientNet-B5 and B, C, D represent DCIM, FPN, and BIT, respectively.

| Setting | | | | LEVIR-CD | | | | WHU-CD | | | | SYSU-CD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | D | IoU | Recall | F1 | Precision | IoU | Recall | F1 | Precision | IoU | Recall | F1 | Precision |
| | | | ✓ | 83.47 | 88.83 | 90.99 | 93.27 | 83.97 | 88.08 | 91.29 | 94.74 | 65.35 | 74.52 | 79.04 | 84.14 |
| ✓ | | | ✓ | 83.64 | 89.18 | 91.09 | 93.09 | 82.36 | 87.02 | 90.33 | 93.9 | 62.34 | 73.21 | 76.8 | 80.77 |
| ✓ | | ✓ | ✓ | 84.44 | 89.43 | 91.56 | 93.81 | 82.42 | 85.51 | 90.36 | 95.8 | 67.9 | 75.86 | 80.88 | 86.62 |
| ✓ | ✓ | ✓ | | 84.8 | 89.7 | 91.78 | **93.95** | **87.17** | 90.36 | **93.15** | **96.1** | 70.77 | **83.68** | 82.88 | 82.1 |
| ✓ | | ✓ | | 84.94 | 90.21 | 91.86 | 93.57 | 84.65 | 88 | 91.68 | 95.68 | 67.04 | 73.82 | 80.27 | **87.96** |
| ✓ | ✓ | ✓ | ✓ | **85.2** | **91.13** | **92.01** | 92.91 | 86.91 | **90.39** | 93 | 95.77 | **71.54** | 80 | **83.41** | 87.13 |

## 4. Conclusion

In this paper, we proposed TriCD, a multi-scale tri-stream interaction network for building change detection from multi-temporal remote sensing imagery. The framework introduces a Difference-guided Cross-temporal Interaction Module (DCIM) for stage-wise feature fusion, a Feature Pyramid Network (FPN) for multi-scale aggregation, and a Transformer-based decoder for global context reasoning. Extensive experiments on three benchmark datasets (LEVIR-CD, WHU-CD, and SYSU-CD) demonstrate that TriCD achieves superior accuracy and structural consistency compared to existing methods.

While the proposed model effectively enhances both local and global representations, its performance remains constrained by the backbone capacity and the current design of feature

interaction. Future research will explore stronger and more flexible interaction modules, adaptive difference feature modeling, and the extension of the framework to larger-scale and multi-class change detection applications.

## References

Bandara, W.G.C., Patel, V.M., 2022. A Transformer-based Siamese Network for Change Detection. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Kuala Lumpur, Malaysia, 207–210. DOI:10.1109/IGARSS46834.2022.9883686.

Chen, H., Shi, Z., 2020. A Spatial–Temporal Attention-based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12(10), 1662. DOI:10.3390/rs12101662.

Chen, H., Qi, Z., Shi, Z., 2022. Remote Sensing Image Change Detection with Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14. DOI:10.1109/TGRS.2021.3095166.

Daudt, R.C., Le Saux, B., Boulch, A., 2018. Fully Convolutional Siamese Networks for Change Detection. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 4063–4067. DOI:10.1109/ICIP.2018.8451652.

Dosovitskiy, A., 2020. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv preprint*, arXiv:2010.11929.

Fang, S., Li, K., Li, Z., 2023. Changer: Feature Interaction is What You Need for Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–11. DOI:10.1109/TGRS.2023.3277496.

Han, C., Wu, C., Guo, H., Hu, M., Li, J., Chen, H., 2023. Change Guiding Network: Incorporating Change Prior to Guide Change Detection in Remote Sensing Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 8395–8407. DOI:10.1109/JSTARS.2023.3310208.

Han, C., Wu, C., Du, B., 2023. HCGMNet: A Hierarchical Change Guiding Map Network for Change Detection. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Pasadena, CA, USA, 5511–5514. DOI:10.1109/IGARSS52108.2023.10283341.

He, P., Shi, W., Zhang, H., et al., 2014. A Novel Dynamic Threshold Method for Unsupervised Change Detection from Remotely Sensed Images. *Remote Sensing Letters*, 5(4), 396–403. DOI:10.1080/2150704X.2014.905656.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. DOI:10.1109/CVPR.2016.90.

Huang, Z., Liang, Q., Yu, Y., Qin, C., 2024. Bilateral Event Mining and Complementary for Event Stream Super-resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 34–43.

Li, K., Cao, X., Meng, D., 2024. A New Learning Paradigm for Foundation Model-based Remote Sensing Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1. DOI:10.1109/TGRS.2024.3365825.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–2125. DOI:10.1109/CVPR.2017.106.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022. DOI:10.1109/ICCV48922.2021.00986.

Lv, Z., Liu, T., Benediktsson, J.A., et al., 2021. Land Cover Change Detection Techniques: Very-high-resolution Optical Images—A Review. *IEEE Geoscience and Remote Sensing Magazine*, 10(1), 44–63. DOI:10.1109/MGRS.2021.3059369.

Munyati, C., 2004. Use of Principal Component Analysis (PCA) of Remote Sensing Images in Wetland Change Detection on the Kafue Flats, Zambia. *Geocarto International*, 19(3), 11–22. DOI:10.1080/10106040408542303.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cham: Springer International Publishing, 234–241. DOI:10.1007/978-3-319-24574-4_28.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint*, arXiv:1409.1556.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the International Conference on Machine Learning (ICML)*, 6105–6114.

Wu, L., Li, Z., Liu, X., et al., 2020. Multi-type Forest Change Detection Using BFAST and Monthly Landsat Time Series for Monitoring Spatiotemporal Dynamics of Forests in Subtropical Wetland. *Remote Sensing*, 12(2), 341. DOI:10.3390/rs12020341.

Zheng, Z., Zhong, Y., Wang, J., et al., 2021. Building Damage Assessment for Rapid Disaster Response with a Deep Object-based Semantic Change Detection Framework: From Natural Disasters to Man-made Disasters. *Remote Sensing of Environment*, 265, 112636. DOI:10.1016/j.rse.2021.112636.