EdgeRefNet: An Edge-Guided Refinement Network for Building Change Detection in Remote Sensing Images

Wafaa I. M. Hussin, Zhi Lu, Aysha Ashraf, Aji Mao, and Zhenming Peng, *Senior Member, IEEE*

***Abstract*—Building change detection (BCD) has achieved remarkable progress with deep learning, yet precise boundary delineation remains a challenging problem. Existing methods often produce blurry contours, incomplete outlines, and building adhesion, partly because structural edge cues are not sufficiently integrated with semantic change representations. In this paper, we propose EdgeRefNet, a hybrid CNN-Transformer architecture for boundary-aware building change detection. The proposed network adopts a dual-path framework consisting of a context path for high-level semantic modeling and an edge path for fine-grained structural detail preservation. Furthermore, a context-guided cross-attention refinement mechanism is introduced to explicitly couple semantic context and edge features, enabling the latter to be refined under semantic guidance for improved boundary localization. To further enhance the edge pathway, we design an Edge Detection Block (EDB) and an Edge Enhancement Module (EEM) to strengthen and refine structural features from shallow representations. Combined with joint supervision on change and edge predictions, EdgeRefNet produces change maps with improved semantic consistency and structural precision. Experiments on the LEVIR-CD and WHU-CD datasets demonstrate that the proposed method outperforms existing approaches, particularly in preserving clearer, more complete, and more accurate building boundaries. The code is available at https://github.com/Wafaa-Hima/EdgeRefNet.**

***Index Terms*—Building change detection, remote sensing images, edge-guided refinement, cross-attention mechanism**

I. Introduction

BUILDING change detection (BCD) aims to identify building-level changes from bi-temporal high-resolution remote sensing images acquired over the same geographic area. As a fundamental task in remote sensing interpretation, BCD is of considerable importance in a wide range of practical applications, including urban expansion analysis, land resource management, illegal construction monitoring, disaster assessment, and post-event damage evaluation [1]–[9].

In recent years, BCD has evolved from traditional change detection methods, such as change vector analysis (CVA) and multivariate alteration detection (MAD), toward deep learning-based approaches. Convolutional neural networks (CNNs) and Vision Transformers (ViTs) have made significant advancements in modeling semantic correspondence and contextual dependencies in bi-temporal imagery, resulting in notable improvements in change detection performance [10]–[16].

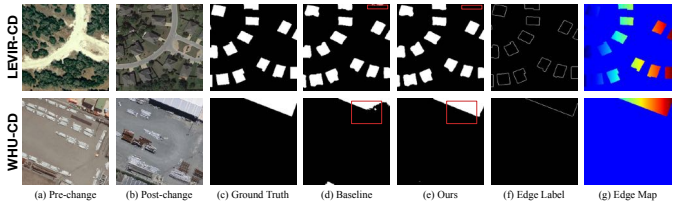


Fig. 1. Visual comparison on the LEVIR-CD and WHU-CD datasets. (a)-(b) bi-temporal images, (c) ground-truth change map, (d) baseline result, (e) our result, (f) edge label, and (g) predicted edge map. Our method yields clearer and more complete boundaries with fewer false positives. Red boxes indicate representative regions.

Despite these advancements, precise boundary delineation remains a non-trivial challenge in BCD. Bi-temporal remote sensing images are often affected by viewpoint variations, illumination differences, seasonal appearance changes, and imperfect registration, which can introduce ambiguity around building contours, particularly in dense urban areas where adjacent structures are closely packed. Under such conditions, methods primarily driven by high-level semantic features may show strong capability in region-level change recognition, while still facing difficulties in preserving fine-grained geometric structures, thereby leading to relatively coarse boundaries, incomplete outlines, building adhesion, and occasional false positives in cluttered environments.

To alleviate this problem, several studies have incorporated edge information into BCD frameworks to enhance structural details and improve boundary localization [17]–[20]. Representative methods, such as EGCTNet [18], demonstrate that edge cues can provide useful structural guidance. However, in many existing edge-guided frameworks, edge information is mainly introduced as an auxiliary signal and is only loosely coupled with semantic representations. As a consequence, edge features contribute more to boundary enhancement than to the learning of change-aware representations, and the consistency between semantic change perception and boundary localization remains limited in challenging scenes. As illustrated in Fig. 1, even when edge information is incorporated, the baseline method [18] may still produce false alarms or incomplete edges in difficult regions. This suggests that introducing edge cues alone may not be sufficient, and that more

Wafaa I. M. Hussin, Aysha Ashraf, Aji Mao, and Zhenming Peng are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mails: wafaaibrahim20@gmail.com; aysha.ashraf92@gmail.com; ajax_mao@163.com; zmpeng@uestc.edu.cn).

Zhi Lu is with the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: zhilu@uestc.edu.cn).

Corresponding author: Zhi Lu.

TABLE I
LIST OF ACRONYMS USED IN THIS PAPER

Acronym	Definition
BCD	Building Change Detection
CNN	Convolutional Neural Network
CVA	Change Vector Analysis
EDB	Edge Detection Block
EEM	Edge Enhancement Module
MAD	Multivariate Alteration Detection
MRFs	Markov Random Fields
PCA	Principal Component Analysis
ViT	Vision Transformer

TABLE II
REPRESENTATIVE BUILDING CHANGE DETECTION METHODS CATEGORIZED BY MODELING PARADIGM AND KEY DESIGN PROPERTIES.

Paradigm	Edge	Representative works
Traditional		CVA [15], [16], MAD [11], PCA [14], MRFs [21], GLRT [22], [23]
CNN	–	FC-Siam [7], STANet [24], DTCDSCN [25], SNUNet [26], TinyCD [27]
	✓	EGRCNN [28], MAEANet [20]
Transformer	–	ChangeFormer [29], STADE-CDNet [1], NATCD [30], STLNet [31],ACWCD [32]
	✓	EATDer [33]
Hybrid	–	BIT [34]
	✓	EGCTNet [18], Edge-CVT [35]

explicit interaction between semantic and edge features could be beneficial for improving boundary-aware change detection.

Motivated by this observation, we propose EdgeRefNet, a hybrid CNN-Transformer architecture for building change detection. The core idea is to strengthen the interaction between high-level semantic information and structural edge cues, so that boundary information can participate more directly in change representation learning. To this end, EdgeRefNet adopts a dual-path design, in which the semantic path is responsible for extracting high-level semantic features and modeling global spatio-temporal dependencies, while the edge path preserves and refines structural details from shallow features.

Building upon this dual-path formulation, we further introduce a context-guided cross-attention refinement mechanism to explicitly couple semantic and edge representations. Specifically, semantic context learned from bi-temporal images is used to guide the refinement of edge features, enabling the model to suppress pseudo-boundaries caused by background textures, illumination variations, and non-change contours while emphasizing genuinely changed building structures. In this way, the proposed framework improves the consistency between change representation and boundary localization, thereby producing structurally more complete and semantically more reliable change maps.

To further enhance the edge pathway, we design two dedicated modules, namely the Edge Detection Block (EDB) and the Edge Enhancement Module (EEM). The EDB is introduced to explicitly strengthen edge-related responses from shallow features, whereas the EEM further refines these responses through channel-spatial enhancement, yielding more robust and discriminative structural representations. The refined edge features are then fused with contextual semantic features in the decoder, and a joint supervision strategy over both change maps and edge maps is employed to further improve the structural quality of the final predictions.

For clarity, the main acronyms used in this paper are summarized in Table I. The main contributions of this work are summarized as follows:

- We propose a hybrid CNN-Transformer architecture for building change detection, which adopts a dual-path context-edge design and a context-guided cross-attention refinement mechanism to improve the interaction between semantic and edge features for boundary-aware change detection.
- We design two dedicated edge modeling modules, namely the Edge Detection Block (EDB) and the Edge Enhancement Module (EEM), to strengthen and refine edge-related structural representations from shallow features.
- Extensive experiments on the LEVIR-CD and WHU-CD datasets demonstrate the effectiveness of the proposed method, especially in preserving clearer, more complete, and more accurate building boundaries.

II. RELATED WORK

The existing literature spans a variety of modeling paradigms, ranging from traditional statistical methods to CNN-based, Transformer-based, and hybrid deep learning models. For clarity, representative approaches are summarized in Table II according to their modeling paradigm and whether explicit edge modeling is incorporated.

A. Traditional Change Detection Methods

Before the widespread adoption of deep learning, change detection mainly relied on statistical modeling and handcrafted features. Representative methods include Change Vector Analysis (CVA) [15], [16], Multivariate Alteration Detection (MAD) [11], and Principal Component Analysis (PCA) [14], which identify changes by modeling pixel-wise differences and applying thresholding or clustering strategies. Probabilistic graphical models such as Markov Random Fields (MRFs) [21] were also introduced to incorporate spatial context into the change inference process. In addition, statistical hypothesis testing methods have shown effectiveness for Synthetic Aperture Radar (SAR) imagery, such as adaptive generalized likelihood ratio tests [22] and multiple covariance equality testing [23]. Although these methods established an important foundation for change detection, they usually depend heavily on manually designed features and have limited robustness in complex urban scenes with illumination variation, shadows, and seasonal appearance changes.

B. Deep Learning-based Change Detection

Deep learning has substantially advanced remote sensing change detection by enabling hierarchical representation learning directly from bi-temporal images. Early studies

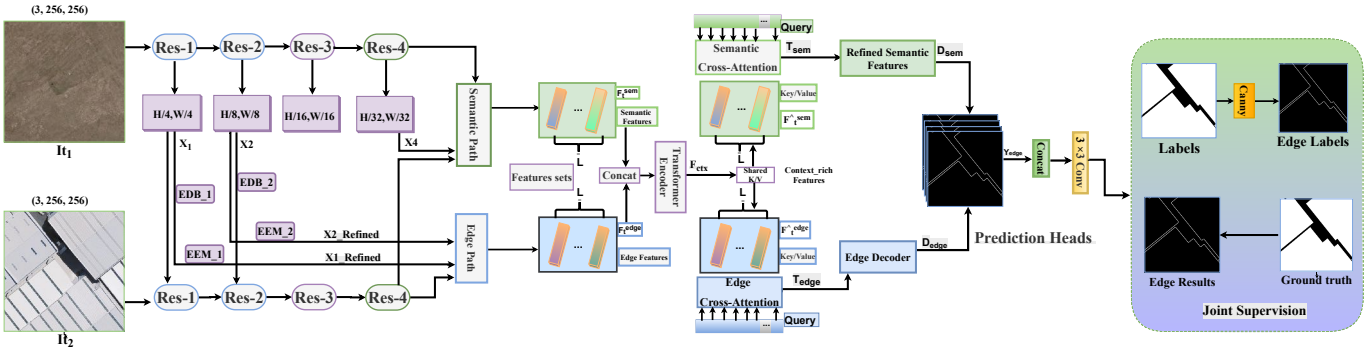


Fig. 2. Overview of the proposed EdgeRefNet architecture. A Siamese ResNet backbone extracts bi-temporal features, which are divided into a semantic path for context modeling and an edge path for structural detail refinement using the proposed EDB and EEM modules. The resulting features are further refined in a context-guided cross-attention decoder, and the network is trained with joint supervision on both change and edge maps.

were predominantly built upon convolutional neural networks (CNNs) [24], [25], in which Siamese architectures became a widely adopted paradigm for comparing features extracted from temporally paired inputs. Representative CNN-based methods include FC-Siam [7], STANet [24], DTCDSCN [36], SNUNet [26], and TinyCD [27]. To improve feature discrimination and spatial detail preservation, these methods commonly incorporate skip connections, multi-scale fusion, and attention mechanisms [28], [36]–[40]. Despite their effectiveness, CNN-based models remain inherently constrained by the locality of convolution, which limits their ability to capture long-range dependencies and broader scene context.

To alleviate this limitation, Transformer-based architectures were introduced into change detection, since self-attention provides a more effective mechanism for modeling long-range interactions and global contextual dependencies. Representative methods in this category include ChangeFormer [29], NATCD [30], STLNet [31], and ACWCD [32]. These studies demonstrate the advantage of Transformer-based representation learning in establishing semantic correspondence across bi-temporal scenes, particularly for complex changes that require broader contextual reasoning.

In parallel, hybrid frameworks have also attracted increasing attention. The motivation behind this line of work is to combine the strong local structural modeling capability of CNNs with the global dependency modeling ability of Transformers. Representative methods such as BIT [34], EGCTNet [18], and Edge-CVT [35] follow this paradigm and show that integrating convolutional inductive bias with self-attention can provide a favorable balance between spatial discrimination and contextual reasoning.

C. Edge-guided Change Detection

Accurate delineation of change boundaries remains a persistent challenge in building change detection. Although high-level semantic features are effective for identifying changed regions, they often lack the spatial precision required to produce sharp and reliable object boundaries. To address this issue, an increasing number of studies have explored boundary-aware and edge-guided designs. [35].

Early edge-guided methods were mainly developed on top of CNN-based frameworks. Some studies introduced edge prediction as an auxiliary task in a multitask learning setting, where structural supervision was used to regularize building extraction or change prediction, while others adopted dedicated edge-aware branches and multi-scale decoding structures to improve boundary localization [19], [41]–[52]. Representative methods in this direction include EGRCNN [28] and MAEANet [20], and edge-guided parallel networks [53], all of which demonstrate that explicit structural cues can improve the sharpness and completeness of predicted change maps.

More recent studies have incorporated attention mechanisms and Transformer components to better exploit global context during boundary refinement. Representative examples include EATDer [33], EGCTNet [18], and Edge-CVT [35]. These methods indicate that edge-aware design can be beneficial for improving boundary localization under more expressive feature interaction schemes. Nevertheless, existing methods still often treat edge information as auxiliary supervision or introduce it through separate branches and relatively shallow fusion. As a result, the interaction between structural edge cues and high-level semantic representations may remain insufficient for fully exploiting their complementary characteristics. This limitation motivates the development of more tightly coupled refinement mechanisms for boundary-aware change detection.

III. Method

A. Problem Formulation

Let $I_{t_1}, I_{t_2} \in \mathbb{R}^{H\times W\times C}$ be a pair of bi-temporal remote sensing images, and let $Y \in \{0,1\}^{H\times W}$ denote the corresponding building change map, where $Y_{i,j} = 1$ indicates that pixel (i,j) belongs to a changed building region, and $Y_{i,j} = 0$ otherwise. The goal of change detection is to learn a function f_θ that predicts a change map $\hat{Y} = f_\theta(I_{t_1}, I_{t_2}) \in \{0,1\}^{H\times W}$. To enhance boundary localization, an auxiliary edge label $Y_{\text{edge}} \in \{0,1\}^{H\times W}$ is further derived from Y using the Canny detector, where $Y_{\text{edge}}(i,j) = 1$ indicates that pixel (i,j) lies on the boundary of a changed building region, and 0 otherwise. The primary task remains the prediction of the change map Y, while Y_{edge} is used as auxiliary supervision during training.

As illustrated in Fig. 2, the proposed network adopts a dual-path architecture consisting of a semantic path and an edge path. The semantic path extracts high-level semantic features to characterize global change patterns, while the edge path preserves fine-grained structural details. A context-guided decoder further leverages semantic context to refine branch features, thereby improving the consistency between semantic change representation and boundary localization.

B. Backbone Feature Extraction

A shared ResNet-18 backbone [54] is used to extract multi-scale features from the bi-temporal input images. For each temporal image I_t, the backbone produces four feature maps corresponding to the outputs of its four residual blocks:

$$X_t^k \in \mathbb{R}^{C_k \times H_k \times W_k}, \quad k \in \{1,2,3,4\}. \tag{1}$$

These feature maps have spatial resolutions of $1/4$, $1/8$, $1/16$, and $1/32$ of the input size, respectively, while the number of channels increases from 64 to 512. Shallow features preserve fine spatial details, whereas deeper features encode higher-level semantic information.

C. Dual-path Feature Extraction

To handle different levels of abstraction, the extracted multi-scale features are divided into two streams. The semantic path operates on the deepest feature map X_t^4 to derive semantic features F_t^{sem}, while the edge path operates on the shallow features X_t^1 and X_t^2 to derive edge features F_t^{edge}.

1) Semantic Path: The semantic path takes the deepest feature map X_t^4 and applies a projection layer to obtain semantic features:

$$F_t^{\text{sem}} = \mathcal{F}_s(X_t^4),\ t \in \{t_1, t_2\}, \tag{2}$$

where $\mathcal{F}_s(\cdot)$ denotes the projection layer implemented by a point-wise convolution.

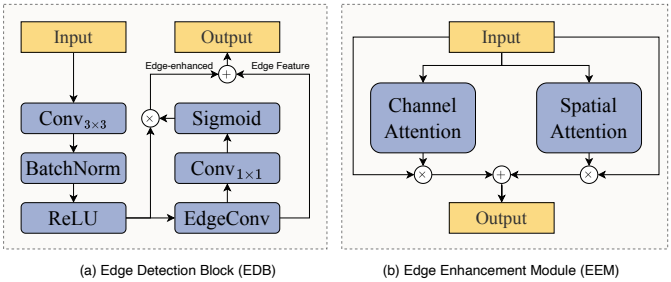
2) Edge Path: The edge path takes the shallow features X_t^1 and X_t^2 as inputs to preserve fine spatial details and structural cues. Each shallow feature is first transformed into an edge-aware representation:

$$E_t^k = \mathcal{G}_e^k(X_t^k),\ k \in \{1,2\}, \tag{3}$$

where $\mathcal{G}_e^k(\cdot)$ denotes the edge feature extraction process for X_t^k, which is detailed in Sec. III-E. Since E_t^2 has a lower spatial resolution than E_t^1, it is upsampled to match the resolution of E_t^1. The two resulting features are then concatenated and fused to generate the final edge representation:

$$F_t^{\text{edge}} = \mathcal{F}_e\left(\text{Concat}(E_t^1, \text{Up}(E_t^2))\right), \tag{4}$$

where $\text{Up}(\cdot)$ denotes the upsampling operation and $\mathcal{F}_e(\cdot)$ denotes the fusion layer.



(a) Edge Detection Block (EDB) (b) Edge Enhancement Module (EEM)

Fig. 3. Illustration of the proposed Edge Detection Block (EDB) and Edge Enhancement Module (EEM) for edge feature extraction and refinement.

D. Context-guided Feature Refinement

Based on the extracted semantic features $F_{t_1}^{\text{sem}}$ and $F_{t_2}^{\text{sem}}$, global spatio-temporal dependencies between the bi-temporal images are modeled by a standard Transformer encoder $\mathcal{T}_{\text{ctx}}$ with positional encoding:

$$F_{\text{ctx}} = \mathcal{T}_{\text{ctx}}\left([F_{t_1}^{\text{sem}}, F_{t_2}^{\text{sem}}]\right), \tag{5}$$

where $[\cdot]$ denotes concatenation. The resulting F_{ctx} serves as a context-aware representation for subsequent feature interaction and refinement.

Based on F_{ctx}, the semantic and edge features are further refined through context-guided cross-attention. Specifically, F_{ctx} serves as the key and value, while the semantic and edge features at each timestamp are used as branch-specific queries:

$$\hat{F}_t^{\text{sem}} = \mathcal{T}_{\text{sem}}(F_t^{\text{sem}}, F_{\text{ctx}}),\ \hat{F}_t^{\text{edge}} = \mathcal{T}_{\text{edge}}(F_t^{\text{edge}}, F_{\text{ctx}}), \tag{6}$$

where $\mathcal{T}_{\text{sem}}(\cdot)$ and $\mathcal{T}_{\text{edge}}(\cdot)$ denote the cross-attention blocks for the semantic and edge branches, respectively.

E. Edge Feature Extraction Modules

To construct the edge-aware representation E_t^k, each shallow feature X_t^k is independently processed by an edge feature extraction module $\mathcal{G}_e(\cdot)$, which is composed of an Edge Detection Block (EDB) followed by an Edge Enhancement Module (EEM), as illustrated in Fig. 3.

For notational simplicity, the internal operations of EDB and EEM are described below using generic input and output symbols, since the same processing is applied independently to each $\{X_t^k\}_{k,t}$.

1) Edge Detection Block (EDB): Given an input feature map H_{in}, the EDB first applies a 3×3 convolution, followed by batch normalization and ReLU activation, to obtain a refined feature representation:

$$H_{\text{feat}} = \text{ReLU}\big(\text{BN}(\text{Conv}_{3\times3}(H_{\text{in}}))\big). \tag{7}$$

An edge feature is then extracted from H_{feat} through an edge-specific convolution:

$$H_{\text{edge}} = \text{EdgeConv}(H_{\text{feat}}). \tag{8}$$

Based on the extracted edge feature, an edge attention map is generated as

$$A_{\text{edge}} = \sigma\big(\text{Conv}_{1\times1}(H_{\text{edge}})\big), \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid activation. The final output of the EDB is obtained by combining the edge-enhanced feature and the extracted edge feature:

$$H_{\text{out}} = H_{\text{feat}} \otimes A_{\text{edge}} + H_{\text{edge}}, \tag{10}$$

where $\otimes$ denotes element-wise multiplication.

2) Edge Enhancement Module (EEM): The feature produced by the EDB is further refined by the EEM. Inspired by the SCSEBlock [42], the EEM applies channel attention and spatial attention in parallel to enhance informative structural responses.

For the channel attention branch, a channel-wise attention vector is computed by global average pooling followed by a shallow MLP and a sigmoid activation:

$$A_{\text{c}} = \sigma\big(\text{MLP}(\text{AvgPool}(H_{\text{in}}))\big). \tag{11}$$

For the spatial attention branch, a spatial attention map is generated using a 1×1 convolution followed by a sigmoid activation:

$$A_{\text{s}} = \sigma\big(\text{Conv}_{1\times1}(H_{\text{in}})\big). \tag{12}$$

The final enhanced feature is obtained by combining the channel-refined and spatially refined features:

$$H_{\text{out}} = H_{\text{in}} \otimes A_{\text{c}} + H_{\text{in}} \otimes A_{\text{s}}. \tag{13}$$

F. Prediction and Learning Objective

After feature refinement, EdgeRefNet performs change prediction and auxiliary edge prediction based on the semantic and edge representations, respectively. Specifically, the bi-temporal feature differences are computed as

$$D_{\text{sem}} = |\hat{F}^{\text{sem}}_{t_1} - \hat{F}^{\text{sem}}_{t_2}|,\ D_{\text{edge}} = |\hat{F}^{\text{edge}}_{t_1} - \hat{F}^{\text{edge}}_{t_2}|. \tag{14}$$

The semantic difference feature D_{sem} is fed into the main prediction head $\mathcal{H}_{\text{main}}$ to produce the pixel-wise change probability map

$$P_{\text{change}} = \mathcal{H}_{\text{main}}(D_{\text{sem}}) \in [0,1]^{H\times W}. \tag{15}$$

Similarly, the edge difference feature D_{edge} is fed into an auxiliary prediction head $\mathcal{H}_{\text{edge}}$ to generate the edge probability map

$$P_{\text{edge}} = \mathcal{H}_{\text{edge}}(D_{\text{edge}}) \in [0,1]^{H\times W}. \tag{16}$$

The network is trained end-to-end with a joint objective that combines the main change detection loss and the auxiliary edge supervision loss, formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cd}} + \lambda \mathcal{L}_{\text{edge}}, \tag{17}$$

where $\mathcal{L}_{\text{cd}}$ denotes the weighted binary cross-entropy loss with respect to the ground-truth change map Y, $\mathcal{L}_{\text{edge}}$ denotes the binary cross-entropy loss imposed on the auxiliary edge label Y_{edge}, and λ balances the contribution of the auxiliary supervision. This design allows the main branch to focus on building change prediction, while the auxiliary edge branch provides additional structural guidance for boundary localization during training.

IV. EXPERIMENTS

A. Datasets

Our experiments were performed on two publicly available datasets for high-resolution building change detection: LEVIR-CD [40] and WHU-CD [25] datasets.

The LEVIR-CD dataset comprises 637 pairs of 1024 × 1024 pixel remote sensing images with a 0.5 m spatial resolution. For our experimental setup, we adhered to the official data split and partitioned the images into non-overlapping patches of 256 × 256 pixels. After this process, we obtained 7,120 training pairs, 1,024 validation pairs, and 2,048 testing pairs.

WHU-CD: This dataset consists of a single pair of very high-resolution aerial images (32,507 × 15,354 pixels) with a spatial resolution of 7.5 cm. As there is no official data split, we first cropped the images into non-overlapping 256 × 256 patches and then randomly split them, yielding 6,096 pairs for training, 762 for validation, and 762 for testing.

B. Baseline Methods

We compare our method with the following baseline change detection methods, including both CNN-based and transformer-based architectures, each offering a distinct approach to change detection.

- **DTCDSCN [36]**: An attention-based method that enhances a Siamese FCN with a dual attention module to capture channel and spatial dependencies.
- **SNUNet [26]**: A multi-scale feature concatenation model combining Siamese networks and Nested-UNet [55] with dense skip connections and channel attention.
- **TinyCD [27]**: A lightweight Siamese U-Net model with a mix and attention mask block for space-semantic attention.
- **BIT [34]**: A transformer-based method that enhances CNN features by modeling long-range dependencies through semantic tokens.
- **ChangeFormer [29]**: A Siamese network with a hierarchical Transformer encoder for extracting features at multiple scales.
- **EGCTNet [18]**: A hybrid CNN-Transformer architecture with an edge-guided branch to refine boundaries and improve change detection accuracy.

C. Evaluation Metrics

We use five standard evaluation metrics for quantitative comparison: precision (P), recall (R), F1-score (F1), intersection over union (IoU), and overall accuracy (OA). Based on the confusion matrix entries true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), these metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{18}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{19}$$

$$F_1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}, \tag{20}$$

TABLE III
COMPARISON OF QUANTITATIVE METRICS FOR LEVIR-CD AND WHU-CD BUILDING CHANGE DETECTION DATASETS

Methods	LEVIR-CD					WHU-CD				
	P (%)	R (%)	F1 (%)	IoU (%)	OA (%)	P (%)	R (%)	F1 (%)	IoU (%)	OA (%)
DTCDSCN [36]	92.83	80.02	85.95	75.37	98.66	63.92	82.30	71.95	56.19	97.42
SNUNet [26]	90.79	89.53	90.15	82.08	99.00	89.00	83.41	86.12	75.62	98.93
TinyCD [27]	92.66	89.52	91.06	83.59	99.10	91.72	**91.76**	91.74	84.74	99.34
BIT [34]	89.24	89.37	89.31	80.68	98.92	86.64	81.48	83.98	72.39	98.75
ChangeFormer [29]	91.22	88.87	90.03	81.87	98.99	90.46	86.84	88.61	79.56	99.11
EGCTNet [18]	90.08	90.56	90.32	82.35	99.01	93.47	88.08	90.70	82.98	99.22
EdgeRefNet	**92.90**	**90.68**	**91.43**	**84.22**	**99.13**	**95.00**	88.81	**91.80**	**84.84**	**99.37**

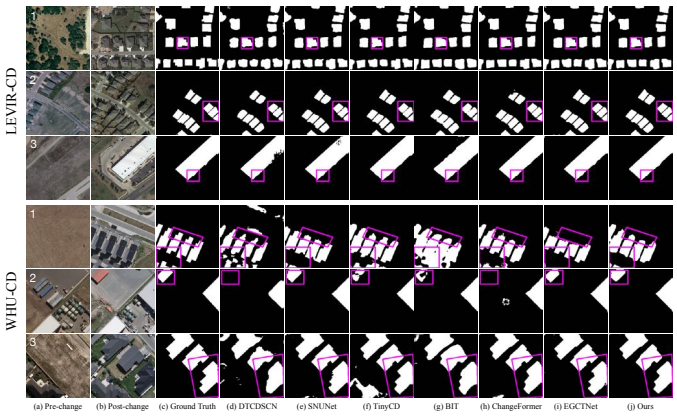


Fig. 4. Qualitative comparisons of different change detection methods. For each dataset, from left to right: (a) pre-change image, (b) post-change image, (c) ground truth, (d) DTCDSCN, (e) SNUNet, (f) TinyCD, (g) BIT, (h) ChangeFormer, (i) EGCTNet, (j) Ours.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{21}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{22}$$

Precision and recall respectively measure the reliability of positive predictions and the detection completeness of changed pixels. The F1-score offers a balanced assessment of these two aspects, while IoU evaluates the spatial alignment between predicted and ground-truth change regions. OA is included for completeness, although it is biased towards the majority no-change class and thus less reliable under class imbalance.

D. Implementation Details

All training and evaluation using a single NVIDIA TITAN RTX 3090 GPU (24 GB) are conducted. In both datasets, the training procedure remained the same. We employed the AdamW optimizer and established the following hyperparameters: an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, a batch size of 8, and a maximum of 200 epochs. A cosine annealing scheduler was utilized to adjust the learning rate during training. To improve the model's generalization and prevent overfitting, the training data was augmented with a series of random transformations, such as resized cropping, 90-degree rotations, color jittering, and Gaussian blurring.

E. Overall Performance

Table III presents the quantitative comparisons on both LEVIR-CD and WHU-CD datasets. Our EdgeRefNet achieves state-of-the-art performance across almost all metrics, validating the effectiveness of explicit edge guidance for change detection.

On LEVIR-CD, EdgeRefNet attains an F1-score of 91.43% and IoU of 84.22%, outperforming all competing methods. Compared to the second-best TinyCD (91.06% F1, 83.59% IoU), our method yields relative improvements of 0.41% in F1 and 0.75% in IoU. TinyCD also surpasses more complex architectures like BIT and ChangeFormer on these key metrics. EGCTNet, which similarly incorporates edge supervision, achieves competitive results with 90.32% F1. On WHU-CD, EdgeRefNet again achieves superior performance with 91.80% F1 and 84.84% IoU. TinyCD closely follows (91.74% F1, 84.74% IoU), while EGCTNet obtains 90.70% F1. These consistent gains across two benchmarks demonstrate the robustness of our edge-guided refinement strategy for building change detection.

The qualitative results, as shown in Fig. 4, reveal common challenges faced by various change detection methods. These challenges include incomplete detection of building parts, boundary adhesion in densely packed small buildings, and irregular false alarms.

For the LEVIR-CD dataset, in the densely packed small buildings (examples 1 and 2), baseline methods are prone to false alarms or the connectivity of boundaries between different buildings. Furthermore, in the third example, the incomplete building detection problem is clearly visible for DTCDSCN and SNUNet. In the WHU-CD dataset, similar issues are observed. In the first example, the baseline methods result in numerous fragmented buildings. In the second example, there is a detection failure of small buildings (DTCDSCN) as well as false alarms (ChangeFormer). In the third example, BIT's boundaries are blurry and not sharp, and all methods exhibit false alarms. In contrast, our method produces clear and complete building detections, effectively addressing these challenges.

F. Ablation Study

1) Effect of Edge Modules: We first conduct ablation studies on the LEVIR-CD dataset to evaluate the effectiveness of EDB and EEM. As shown in Table IV, the baseline model achieves an F1-score of 91.23%, with the highest precision

TABLE IV
ABLATION EXPERIMENTS ON THE LEVIR-CD DATASET

Methods	P (%)	R (%)	F1 (%)	IoU (%)	OA (%)
Baseline	**93.26**	89.29	91.23	83.88	99.12
w/ EDB	92.81	89.71	91.23	83.88	99.12
w/ EEM	92.15	90.63	91.38	84.13	**99.13**
Full Model	92.90	**90.68**	**91.43**	**84.22**	**99.13**

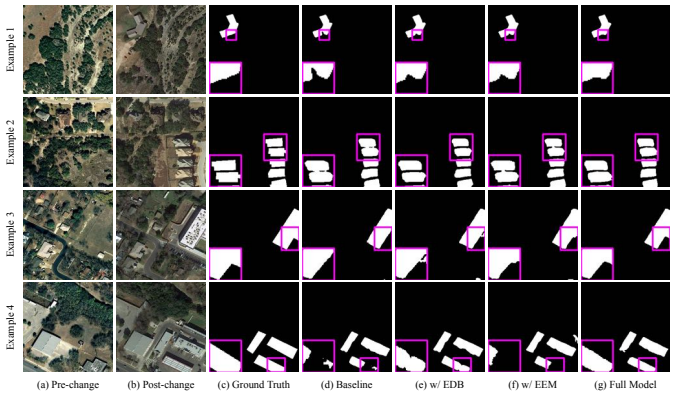


Fig. 5. Qualitative results of the ablation study on the LEVIR-CD dataset, with zoomed-in patches in the lower-left corners highlighting the boundary refinement achieved by the proposed modules.

of 93.26% and the lowest recall of 89.29%. This suggests that, in the absence of edge assistance, the model tends to be conservative in predicting positive classes, which may result in under-segmented predictions where boundaries contract inward, leaving the outer edges of objects undetected. Introducing edge-related modules significantly improved recall, but at the cost of precision, as enhancing the edges also introduced some false positives.

In terms of overall performance, adding EDB alone results in performance nearly identical to the baseline. Although EDB provides explicit edge priors to guide the network toward boundary regions, without subsequent feature refinement, these priors may introduce additional false alarms. Adding EEM alone improves performance through its channel-spatial attention, which amplifies informative signals, but it may also lead to unfocused enhancement. The full model integrates EDB and EEM, combining the spatial localization from EDB with the feature refinement from EEM, allowing the model to achieve more localized enhancement while overcoming the limitations of each module when used independently.

We further visualize different examples in Fig. 5 to highlight the differences. Overall, without boundary guidance, the baseline tends to cause boundary inward contraction (example 1), merging of nearby boundaries (example 2), missing parts of buildings (example 3), or the segmentation of a building into multiple parts (example 4). Furthermore, EDB, through explicit edge guidance, significantly preserves boundary information but is prone to false alarms. Using EEM alone mitigates the boundary issues of the baseline, but it doesn't achieve the same clarity as EDB. Combining both modules produces the best results, which further emphasizes the importance of edge guidance in improving boundary detection and localization in challenging scenarios.

TABLE V
ABLATION ON INPUT CHANNELS ON LEVIR-CD

Input	P (%)	R (%)	F1 (%)	IoU (%)	OA (%)
Grayscale	90.03	83.67	86.73	76.57	98.69
RGB	**92.90**	**90.68**	**91.43**	**84.22**	**99.13**

TABLE VI
EFFICIENCY COMPARISON OF DIFFERENT METHODS

Model	Params. (M)	FLOPs (G)	Inference Time (ms)	
			LEVIR-CD	WHU-CD
DTCDSCN [36]	41.07	13.22	21.23 $\pm$ 0.48	21.56 $\pm$ 0.48
SNUNet [26]	12.03	54.83	21.18 $\pm$ 0.48	21.18 $\pm$ 0.60
TinyCD [27]	0.29	1.54	15.34 $\pm$ 0.35	16.05 $\pm$ 1.67
BIT [34]	11.99	4.90	19.46 $\pm$ 0.24	19.53 $\pm$ 0.32
ChangeFormer [29]	41.03	202.79	56.39 $\pm$ 3.15	54.16 $\pm$ 0.51
EGCTNet [18]	106.16	38.47	134.11 $\pm$ 2.13	132.88 $\pm$ 2.53
EdgeRefNet	19.48	27.55	61.17 $\pm$ 2.50	65.35 $\pm$ 5.83

2) Effect of Input Channels: To evaluate the influence of input channel configuration, we compare the default 3-channel RGB input with a single-channel grayscale variant on the LEVIR-CD dataset. As shown in Table V, the RGB input consistently achieves better performance than the grayscale setting across all evaluation metrics. In particular, the F1 score improves from 86.73% to 91.43%, and the IoU increases from 76.57% to 84.22%. These results indicate that RGB input preserves richer appearance information across the visible bands than the grayscale.

G. Computational Efficiency

To evaluate the computational efficiency of the proposed method, we compare its parameter count, FLOPs, and GPU inference time with several representative change detection models on the LEVIR-CD and WHU-CD datasets. The results are reported in Table VI. All inference-time measurements are obtained on a server equipped with an NVIDIA RTX 3090 GPU and are reported as the average runtime per image pair.

As shown in Table VI, lightweight CNN-based models such as TinyCD, BIT, SNUNet, and DTCDSCN require less computation and lower inference time than the heavier architectures. TinyCD is the fastest method on both datasets, whereas ChangeFormer and EGCTNet require noticeably more computation.

EdgeRefNet contains 19.48 M parameters and 27.55 GFLOPs, with inference times of 61.17 $\pm$ 2.50 ms on LEVIR-CD and 65.35 $\pm$ 5.83 ms on WHU-CD. Although it is slower than the lightweight baselines, it is clearly more efficient than EGCTNet and requires substantially fewer FLOPs than ChangeFormer. Taken together with the quantitative results, these comparisons indicate that EdgeRefNet achieves a reasonable trade-off between detection accuracy and computational cost.

H. Visualization Results

To provide a more intuitive visualization of the proposed method, Fig. 6 presents the predicted change maps and the

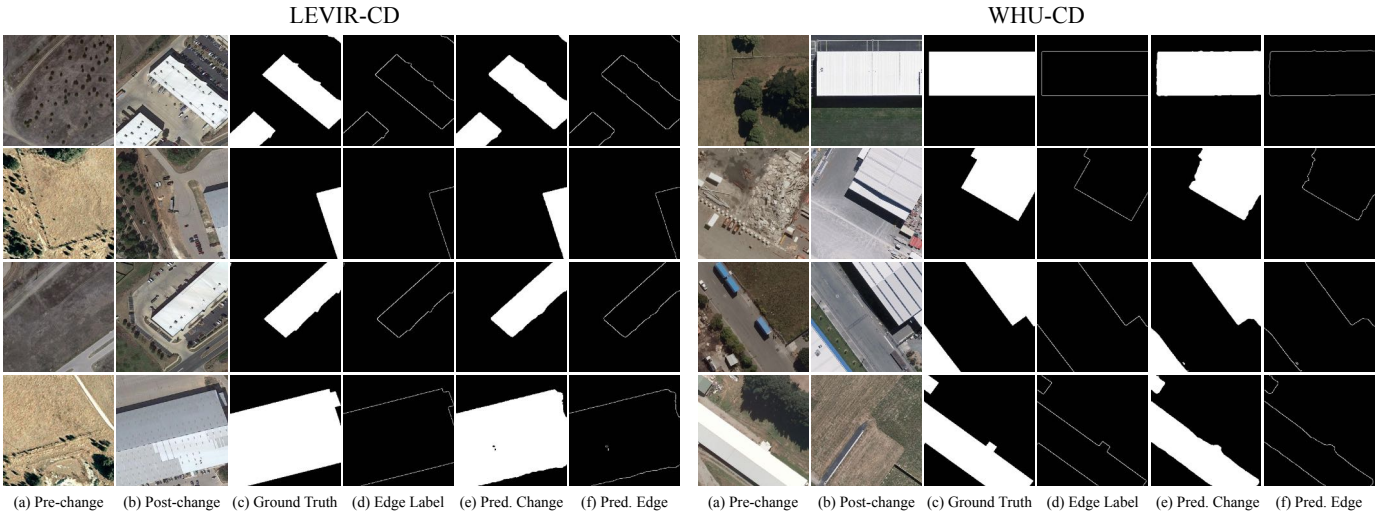


Fig. 6. Qualitative results of the proposed method on the LEVIR-CD and WHU-CD datasets. The figure demonstrates our model's ability to accurately predict building change maps while maintaining high-fidelity structural edges. (a)-(b) Bi-temporal images, (c)-(d) Ground truth change map and edges, (e)-(f) Predicted change map and edges.

Fig. 7. Visual results of 256 × 256 images on the LEVIR-CD dataset. (a) pre-change image, (b) post-change image, (c) ground truth, (d) Change map produced by EdgeRefNet.

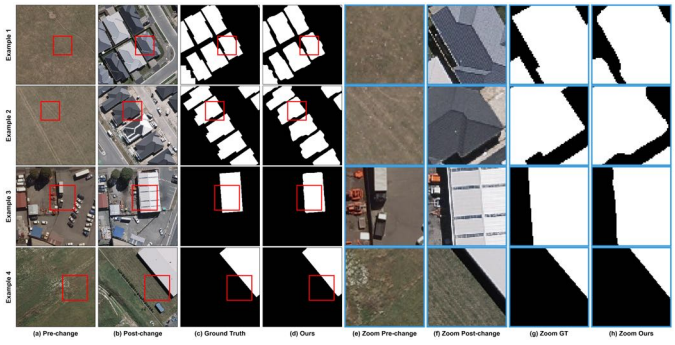


Fig. 8. Visual results of 256 × 256 images on the WHU-CD dataset. (a) pre-change image, (b) post-change image, (c) ground truth, (d) change map produced by EdgeRefNet.

corresponding edge maps on the LEVIR-CD and WHU-CD datasets, while Figs. 7 and 8 further provide high-resolution zoomed-in results in several complex scenarios. It can be observed that the predicted edges generally align well with the boundaries of the changed buildings, and the associated change maps preserve relatively clear and complete building structures. The zoomed-in regions further show that EdgeRefNet maintains good geometric integrity and regional completeness across buildings with different scales, shapes, and spatial distributions. These visual results complement the quantitative analysis by providing an intuitive view of the predictions produced by the proposed method.

V. Conclusion

In this paper, we proposed EdgeRefNet, a hybrid CNN-Transformer framework for building change detection that explicitly enhances the interaction between semantic and edge-aware representations. By introducing the Edge Detection Block (EDB), the Edge Enhancement Module (EEM), and the cross-attention-based refinement mechanism, the proposed method improves the delineation of changed building regions, especially around complex boundaries. Extensive experiments on the LEVIR-CD and WHU-CD datasets demonstrate that EdgeRefNet achieves strong quantitative performance and produces clear and complete change maps. Although the proposed framework effectively strengthens semantic-edge interaction, it does not explicitly exploit the geometric regularity of buildings. Future work will explore the incorporation of geometry-aware structural constraints to further improve the preservation of shape consistency, boundary quality, and region integrity in complex urban scenes.

Acknowledgment

This work was supported by the Natural Science Foundation of Sichuan Province, China, under Grant 2025ZNSFSC0522, and by the National Natural Science Foundation of China under Grant 61571096. The authors would like to thank Dr. Nihad A. A. Elhag and Dr. Mustafa O. Ali for their valuable discussions and suggestions.

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 Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[2] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021.

[3] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[4] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal vhr images based on deep kernel pca convolutional mapping network," *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 12084–12098, 2021.

[5] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[6] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, "Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2020.

[7] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE international conference on image processing (ICIP)*, pp. 4063–4067, IEEE, 2018.

[8] Y. Zhong, W. Liu, J. Zhao, and L. Zhang, "Change detection based on pulse-coupled neural networks and the nmi feature for high spatial resolution remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 537–541, 2014.

[9] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *International journal of remote sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[10] M. J. Canty, *Image analysis, classification and change detection in remote sensing: with algorithms for Python*. Crc Press, 2019.

[11] B. Wang, S.-K. Choi, Y.-K. Han, S.-K. Lee, and J.-W. Choi, "Application of ir-mad using synthetically fused images for change detection in hyperspectral data," *Remote sensing letters*, vol. 6, no. 8, pp. 578–586, 2015.

[12] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS Journal of photogrammetry and remote sensing*, vol. 80, pp. 91–106, 2013.

[13] Y.-Q. Cheng, H.-C. Li, T. Celik, and F. Zhang, "Frft-based improved algorithm of unsupervised change detection in sar images via pca and k-means clustering," in *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS*, pp. 1952–1955, IEEE, 2013.

[14] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE geoscience and remote sensing letters*, vol. 6, no. 4, pp. 772–776, 2009.

[15] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2006.

[16] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with landsat," in *LARS symposia*, p. 385, 1980.

[17] S. Holail, T. Saleh, X. Xiao, M. Zahran, G.-S. Xia, and D. Li, "Edge-CVT: Edge-informed CNN and vision transformer for building change detection in satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 227, pp. 48–68, 2025.

[18] L. Xia, J. Chen, J. Luo, J. Zhang, D. Yang, and Z. Shen, "Building change detection based on an edge-guided convolutional neural network combined with a transformer," *Remote Sensing*, vol. 14, no. 18, p. 4524, 2022.

[19] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[20] B. Yang, Y. Huang, X. Su, and H. Guo, "Maeanet: Multiscale attention and edge-aware siamese network for building change detection in high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 19, p. 4895, 2022.

[21] T. Kasetkasem and P. Varshney, "An image change detection algorithm based on markov random field models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 8, pp. 1815–1823, 2002.

[22] H. Zhuang, Z. Tan, K. Deng, and G. Yao, "Adaptive generalized likelihood ratio test for change detection in sar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 416–420, 2020.

[23] D. Ciuonzo, V. Carotenuto, and A. De Maio, "On multiple covariance equality testing with application to sar change detection," *IEEE Transactions on Signal Processing*, vol. 65, no. 19, pp. 5078–5091, 2017.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

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

[26] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[27] A. Codegoni, G. Lombardi, and A. Ferrari, "Tinycd: A (not so) deep learning model for change detection," *Neural Computing and Applications*, vol. 35, no. 11, pp. 8471–8486, 2023.

[28] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[29] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 207–210, IEEE, 2022.

[30] Z. Guo, H. Chen, and F. He, "NATCD: A Multi-Scale Neighborhood Attention Transformer Network for Remote Sensing Image Change Detection," in *2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 10311–10314, 2024.

[31] L. Mei, A. Huang, Z. Ye, Y. Yalikun, Y. Wang, C. Xu, W. Yang, and X. Li, "STLNet: Symmetric transformer learning network for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 2655–2667, 2025.

[32] W. Liu, Z. Yu, and B. Luo, "ACWCD: Utilizing inherent transformers information and prior knowledge for weakly supervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025.

[33] J. Ma, J. Duan, X. Tang, X. Zhang, and L. Jiao, "EATDer: Edge-Assisted Adaptive Transformer Detector for Remote Sensing Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[34] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[35] S. Holail, T. Saleh, X. Xiao, M. Zahran, G.-S. Xia, and D. Li, "Edge-CVT: Edge-informed CNN and vision transformer for building change detection in satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 227, pp. 48–68, 2025.

[36] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[37] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2020.

[38] 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 Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[39] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *International workshop on deep learning in medical image analysis*, pp. 3–11, Springer, 2018.

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

[41] G. Yang, Q. Zhang, and G. Zhang, "Eanet: Edge-aware network for the extraction of buildings from aerial images," *Remote Sensing*, vol. 12, no. 13, p. 2161, 2020.

[42] T. Liu, J. Li, W. Cao, M. Tang, and G. Yang, "Mlcnet: Multitask level-specific constraint network for building change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 11823–11838, 2024.

[43] J. Zhang, Z. Shao, Q. Ding, X. Huang, Y. Wang, X. Zhou, and D. Li, "Aernet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[44] A. Eftekhari, F. Samadzadegan, and F. D. Javan, "Building change detection using the parallel spatial-channel attention block and edge-guided deep network," *International Journal of Applied Earth Observation and Geoinformation*, vol. 117, p. 103180, 2023.

[45] X. Li, L. Xie, C. Wang, J. Miao, H. Shen, and L. Zhang, "Boundary-enhanced dual-stream network for semantic segmentation of high-resolution remote sensing images," *GIScience & Remote Sensing*, vol. 61, no. 1, p. 2356355, 2024.

[46] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[47] J. Lin, G. Wang, D. Peng, and H. Guan, "Edge-guided multi-scale foreground attention network for change detection in high resolution remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 133, p. 104070, 2024.

[48] Y. Liu, F. Liu, J. Liu, and L. Xiao, "Edge-object co-driven learning for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[49] P. Wang, F. Cheng, Y. Yao, L. Liu, J. Zhang, A. Bouras, D. Narasimhan, L. Qin, S. Wang, and C. Liu, "Enhancing change detection with edge-guided difference modeling in remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, 2025.

[50] C. You, N. Wang, D. Zhu, R. Liu, and W. Li, "High-resolution remote sensing change detection with edge-guided feature enhancement," *IEEE Geoscience and Remote Sensing Letters*, 2025.

[51] Y. Xing, J. Hu, Y. Jia, and R. Huang, "Multi-scale edge enhancement and progressive change-aware network for remote sensing change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

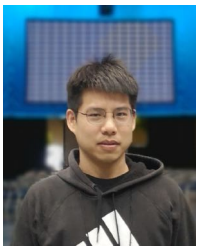
[52] M. Li, D. Ming, L. Xu, D. Dong, and Y. Zhang, "Sfearnet: A network combining semantic flow and edge-aware refinement for highly efficient remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[53] Y. Zhu, K. Lv, Y. Yu, and W. Xu, "Edge-guided parallel network for vhr remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 7791–7803, 2023.

[54] 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.

[55] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

Wafaa I. M. Hussin received the M.S. degree in computer architecture and networking from the Faculty of Engineering, University of Khartoum, Khartoum, Sudan, in 2019. She is currently pursuing the Ph.D. degree in information and communication engineering with the University of Electronic Science and Technology of China (UESTC), Chengdu, China. She also works as a researcher at the National Energy Research Center, Ministry of Higher Education and Scientific Research, Khartoum, Sudan. Her research interests include image processing, computer vision, and change detection.



Zhi Lu received the B.S., M.Phil., and Ph.D. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2015, 2018, and 2022, respectively. He is currently Associate Professor at the Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China. His research interests include multimedia content understanding and computer vision.

Aysha Ashraf received the Master's degree in computer science and is currently pursuing the Ph.D. degree in computer science at the University of Electronic Science and Technology of China (UESTC), Chengdu. She has over five years of experience as a lecturer in computer science.

Her research interests include remote sensing, computer vision, AI-driven image processing, and multimodal event-based change detection using vision–language models.

Aji Mao received the B.S. degree in electronic information engineering from China University of Mining and Technology (CUMT), Xuzhou, China, in 2021. He is currently pursuing the M.S. degree with the School of Information and Communication Engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China.

His research interests include computer vision, large language models, and infrared target recognition.

Zhenming Peng (Senior Member, IEEE) received the Ph.D. degree in geodetection and information technology from the Chengdu University of Technology, Chengdu, China, in 2001.

From 2001 to 2003, he was a postdoctoral researcher with the Electronics Chinese Academy of Sciences, Chengdu. He is currently a professor at the University of Electronic Science and Technology of China, Chengdu. His research interests include image processing, signal processing, and target recognition and tracking.