Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Scale Temporal Difference Refinement for Building Change Detection in High- Resolution Remote Sensing Images

Hao Yu
Jilin Jianzhu University
Shenzhao Geng
Jilin Jianzhu University
Nan Lin
Modern Industry College, Jilin Jianzhu University
Yuqiang Zhang
zhangyuqiang@jlju.edu.cn

Modern Industry College, Jilin Jianzhu University
Chuanpeng Zhao
National Forestry and Grassland Administration, Chinese Academy of Sciences
Huiying Li
Qingdao University of Technology

Article

Keywords: High-resolution remote sensing images, Building change detection, Transformer, Multi-scale modeling

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11685242/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multi-Scale Temporal Difference Refinement for Building Change Detection in High-Resolution Remote Sensing Images

Hao Yu, Shenzhao Geng, Nan Lin, Yuqiang Zhang, Chuanpeng Zhao, Huiying Li

Abstract— Transformer architectures enable change detectors to exploit long-range spatial context, but their predictions may still be disturbed by background variations, scale diversity, and acquisition differences between two temporal images. We develop TDBIT by extending the Bitemporal Image Transformer (BIT) with an explicit Temporal Difference Module (TDM). After contextual information has been injected into the two feature maps by the Transformer decoder, TDM computes their absolute discrepancy and processes it through three parallel dilated-convolution branches. The branches use dilation rates of 1, 3, and 5 to characterize change responses over different spatial extents. Their outputs are fused with the original discrepancy, allowing the network to combine fine-scale discrepancy patterns with wider contextual evidence. A training objective combining cross-entropy and focal binary cross-entropy (Focal-BCE) is further employed to reduce the relative contribution of easily classified pixels and place greater emphasis on difficult samples. On LEVIR-CD and WHU-CD, TDBIT obtains F1-scores of 90.34% and 89.65%, together with IoU values of 82.38% and 81.24%, respectively. These results are higher than those of the BIT baseline on the two evaluated datasets, while the ablation results support the contributions of TDM and the combined objective under the current experimental settings.

Index Terms—High-resolution remote sensing images, Building change detection, Transformer, Multi-scale modeling

Hao Yu is with the School of Geomatics and Prospecting Engineering, Jilin Jianzhu University, Changchun 130118, China (e-mail: yuhao@jlju.edu.cn).
Shenzhao Geng is with the School of Geomatics and Prospecting Engineering, Jilin Jianzhu University, Changchun 130118, China (e-mail: 13131125685@163.com).
Nan Lin is with the Modern Industry College, Jilin Jianzhu University, Changchun 130118, China (e-mail: linnan@jlju.edu.cn).
Yuqiang Zhang is with the Modern Industry College, Jilin Jianzhu University, Changchun 130118, China, and is the corresponding author (e-mail: zhangyuqiang@jlju.edu.cn).
Chuanpeng Zhao is with the Key Laboratory of Marsh Wetland Ecosystem Conservation and Restoration, National Forestry and Grassland Administration, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China (e-mail: zhaochuanpeng@iga.ac.cn).
Huiying Li is with the School of Environmental and Municipal Engineering, Qingdao University of Technology, Qingdao 266520, China (e-mail: lihy@qut.edu.cn).

I. INTRODUCTION

Remote-sensing change detection compares observations of the same location acquired at different times and produces a spatial map of the regions that have changed. The task supports applications such as urban expansion analysis, map updating, disaster investigation, and land management [1]-[4]. Building change detection is particularly relevant because the appearance, removal, or modification of buildings provides direct evidence of urban development. Its performance is nevertheless affected by non-change factors, including lighting conditions, seasonal differences, shadows, viewing geometry, and registration errors. Variations in building size and shape further complicate the delineation of small targets and narrow boundaries. A practical detector must therefore separate genuine structural changes from temporal appearance variations while retaining sufficient spatial detail.

Early change detection approaches relied on manually designed descriptors and statistical comparisons, such as image differencing, change-vector analysis, principal-component analysis, slow-feature analysis, and object-based techniques [1], [2], [7]-[9]. Their interpretability is useful, but their results are strongly influenced by the selected features, decision thresholds, and image quality. Learning-based methods reduce this dependence by deriving temporal representations from annotated image pairs [3], [6], [10]. Fully convolutional and Siamese architectures in particular encode the two observations and estimate change evidence from concatenated, subtracted, or otherwise interacted features [11], [12].

CNN-based detectors have subsequently incorporated encoder-decoder recovery, cross-level fusion, attention-based recalibration, and deep supervision [13]-[20]. These additions improve semantic representation and help restore details after feature downsampling. However, the basic convolution operator aggregates information within a limited neighborhood. Dependencies between distant regions must consequently be built through multiple layers or enlarged kernels, and relationships between the two acquisition times are often learned only through indirect feature interaction.

Self-attention has been adopted in remote-sensing change detection to establish interactions between spatially distant locations and to provide a broader contextual view [21]-[25]. BIT reduces the cost of global modeling by summarizing each high-resolution feature map with a small number of semantic tokens. The tokens from the two dates are jointly updated by a Transformer and are then returned to the spatial domain for dense prediction [25]. This token-based design retains the convolutional encoder while avoiding attention over the entire pixel grid. Later studies have explored hierarchical Transformer

backbones, Siamese Transformer structures, and explicit temporal interaction mechanisms for change detection [29]-[35].

Despite the global interaction provided by BIT, its final change representation is obtained mainly by comparing the two decoded features directly. Such a comparison does not explicitly separate discrepancy patterns according to their spatial extent. Broad construction or demolition areas require contextual support over a relatively large region, whereas small buildings, partial alterations, and thin boundaries depend on fine local responses. Shadows, texture changes, and minor alignment errors can also create local differences that resemble real changes. These factors may lead to incomplete structures or spurious activations in the difference map generated by BIT.

Motivated by this issue, we propose TDBIT, a BIT-based detector with explicit multi-scale temporal difference refinement. The original semantic tokenization and Transformer-based global interaction are retained. After the decoder, a Temporal Difference Module computes the absolute discrepancy between the two enhanced feature maps and processes it with three dilated-convolution branches. The branches use different receptive fields and are combined with the initial discrepancy before classification. This arrangement allows the change representation to include both localized boundary evidence and wider structural context. In addition, the training objective combines cross-entropy with weighted Focal-BCE to reduce the dominance of easy samples during optimization.

TDBIT is evaluated on the public LEVIR-CD and WHU-CD building change detection datasets. Quantitative comparisons, visual predictions, and ablation experiments are reported to examine the performance of the proposed module and the combined objective.

II. RELATED WORK

A. CNN-Based Remote Sensing Image Change Detection Methods

Convolutional change detectors differ primarily in the point at which information from the two dates is allowed to interact. Early-fusion models stack the image pair before feature extraction, whereas Siamese models encode each observation separately with shared parameters and combine the resulting features through concatenation, subtraction, or learned interaction [11], [12]. Feature-level comparison has become particularly common because it separates temporal representation learning from the subsequent estimation of change evidence.

Later CNN architectures improved this basic pipeline through encoder-decoder reconstruction, multi-level feature aggregation, deep supervision, and attention-based refinement [36]–[41]. For remote sensing imagery, these mechanisms have also been combined with boundary constraints, cross-level interaction, difference enhancement, and local-global context aggregation [13]–[20], [42]–[48], [53]. They strengthen the recovery of object details and help localize narrow change boundaries. Nevertheless, convolutional interaction is accumulated through spatially bounded kernels. Modeling relationships between distant regions can therefore require deep feature hierarchies, and similar background patterns may still generate ambiguous responses in dense urban scenes.

B. Transformer-Based Remote Sensing Image Change Detection Methods

Self-attention provides an alternative means of relating spatially separated regions and corresponding information from two acquisition times. One line of research first compresses high-resolution feature maps into a compact token representation. BIT follows this strategy and performs global bitemporal interaction in token space before restoring the contextual information to the pixel-level features [25]. This avoids applying global attention directly to every spatial location.

Other methods retain hierarchical spatial representations and use Siamese Transformer encoders, pyramidal processing, or window-based attention to exchange information across multiple resolutions [29], [30], [49]–[51]. A further group introduces explicit cross-temporal attention or difference-aware interaction so that temporal correspondence is learned directly rather than inferred only from independently enhanced features [31], [32]. Although these designs improve global reasoning, fine structures can be weakened during tokenization or hierarchical abstraction. For building change detection, preserving small objects, irregular contours, and narrow boundary responses therefore remains an important complement to global context modeling [33]–[35].

C. BIT Network and Its Limitations

BIT is selected as the baseline because its semantic-token mechanism offers scene-level interaction at substantially lower cost than full-resolution self-attention [25]. A shared convolutional encoder first extracts the two temporal feature maps. The tokenizer summarizes each map with a small set of semantic tokens, and a Transformer models their bitemporal relationships. The decoder subsequently returns the contextualized token information to the spatial feature maps.

The remaining limitation lies mainly in the construction of the final change representation. Direct differencing provides a useful first estimate, but it does not deliberately distinguish discrepancy patterns associated with small buildings, broad structures, or narrow boundaries. Moreover, changed pixels generally occupy a much smaller area than unchanged pixels, so standard optimization can be dominated by easy background samples. TDBIT addresses these two issues through multi-scale refinement of the decoded feature difference and an objective that assigns greater influence to difficult samples.

III. METHODOLOGY

A. Overall Network Architecture

The processing sequence of TDBIT is shown in Fig. 1. The two temporal observations are encoded by the same ResNet-18 to obtain features in a shared representation space. Each feature map is then summarized by a semantic tokenizer. The token sequences from the two dates are joined and updated by the Transformer encoder, after which the decoder maps the contextual information back to the spatial domain. TDM receives the two decoded maps, calculates their absolute difference, and performs multi-scale convolutional refinement. The resulting feature is sent to the prediction head, which produces logits for the unchanged and changed classes at every pixel.

For the experiments, BIT is instantiated using the base_transformer_pos_s4_dd8 configuration. Its tokenization module produces four semantic vectors, each represented in a 32-dimensional space. The network includes a single encoder

layer and eight decoder layers, with dropout disabled throughout the Transformer. Each attention module uses eight heads, and every head operates on 64 features. Thus, attention is actually performed in a 512-dimensional latent space, despite the 32-dimensional representation at the module interface. Specifically, a learnable projection expands each incoming token from 32 to 8 × 64features before the projected representation is partitioned across the eight heads. Once attention has been computed, the head outputs are reassembled and compressed to 32 dimensions by the final projection. Consequently, the attention module can employ a wider internal representation without altering the dimensionality of the tokens passed between Transformer sub-layers.

Within TDM, three dilated convolution paths operate in parallel to capture difference information at multiple receptive-field scales. Each path generates a 16-channel feature map. The outputs of these paths are stacked along the channel axis, and a pointwise convolution subsequently transforms the aggregated tensor into a 32-channel representation. This multi-scale result is then combined with the initial difference map to obtain a refined feature for change recognition. For dense prediction, the refined feature is spatially enlarged by a factor of four and processed by a classifier composed of two convolutional layers. The classifier assigns two scores to every pixel, representing the change and no-change categories.

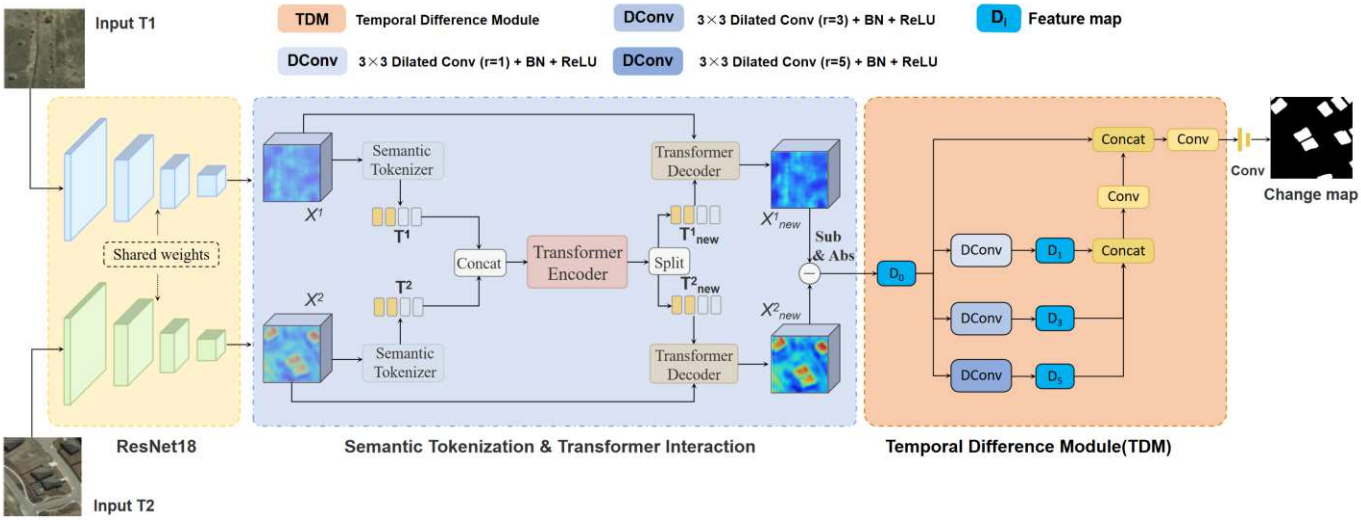


Fig. 1. Schematic representation of the TDBIT framework.

1) Shared-Backbone Feature Encoding: A single ResNet-18 [52] is employed to independently encode the paired observations, with its weights reused for both temporal branches. The backbone follows the output-stride adjustment adopted in BIT [25]. Specifically, the first residual units in stages 3 and 4 operate with a stride of 1 rather than 2, and the projection shortcuts are configured accordingly. Eliminating these two downsampling steps increases the resolution of the deep features, which is beneficial for identifying small targets and accurately locating change boundaries. The backbone output is subsequently compressed to 32 channels using a pointwise convolution and then upsampled to twice its spatial resolution. Therefore, an input of 256×256 yields a $32 \times 64 \times 64$ feature tensor for each observation.

Denote the paired input images by $\mathbf{I}^1$ and $\mathbf{I}^2$. Their feature extraction process can be expressed as

$$X^i = f_b(I^i), \quad i \in \{1,2\} \tag{1}$$

where $f_b(\cdot)$ represents the ResNet-18 encoder shared by the two branches, and $X^i \in R^{C\times H\times W}$ is the representation associated with the i-th observation. In the present implementation, $C = 32$ and $H = W = 64$ for a 256×256 image patch. This relatively dense feature grid retains the spatial cues needed to distinguish fine building contours and small changed regions while still providing high-level semantic information. Before semantic token generation, the two spatial axes of X^i are collapsed, converting the tensor into a sequence of HW feature vectors.

2) Semantic Token Construction: Directly performing global self-attention on high-resolution feature maps introduces a high computational cost. To address this issue, this study adopts a Siamese Semantic Tokenizer with shared parameters to compress the bitemporal convolutional features into a small number of semantic tokens [25]. Let the flattened feature of the i-th temporal phase be denoted as $X^i \in R^{HW\times C}$, and the corresponding token set be denoted as $T^i \in R^{L\times C}$, where L represents the number of semantic tokens.

For the i-th temporal phase, L semantic grouping maps are first generated through a point-wise linear projection. A softmax operation is then performed along the spatial dimension HW to obtain normalized spatial attention weights. Finally, the original features are weighted and aggregated according to these attention weights to produce semantic tokens. This process can be formulated as

$$T^i = (A^i)^T X^i = \left(\mathrm{Softmax}\left(\phi(X^i; W)\right)\right)^T X^i \tag{2}$$

where $\phi(\cdot;W)$ denotes the point-wise linear projection, $W \in R^{C\times L}$ is a learnable parameter matrix, and $A^i \in R^{HW\times L}$ represents the spatial attention weight matrix obtained by softmax normalization. Therefore, the l-th token corresponds to a global spatial attention distribution over the entire feature map. The final token set $T^i \in R^{L\times C}$ serves as a compact semantic representation of the i-th temporal phase.

Fig. 2 illustrates the semantic token construction process for a single temporal phase. For bitemporal features, the same operation is applied separately to the two temporal phases using a shared-parameter Semantic Tokenizer. This process compresses the original high-resolution features into a small number of compact semantic representations, thereby reducing the

computational complexity of the Transformer while preserving high-level semantic information in the bitemporal images.

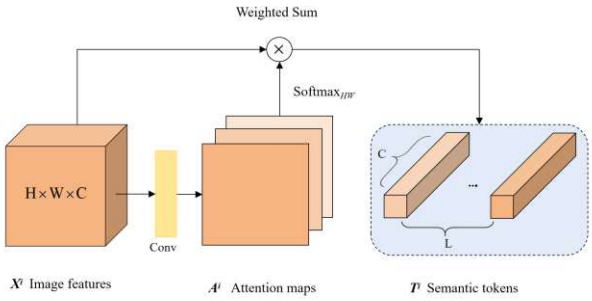


Fig. 2. Schematic illustration of the semantic tokenization module.

3) Transformer Encoder and Decoder: After obtaining the bitemporal semantic tokens, the tokens from the two temporal phases are concatenated along the sequence dimension. A learnable positional encoding is then added to preserve positional information in the token sequence [25]:

$$T = \mathrm{Concat}(T^1, T^2) + W_{PE} \tag{3}$$

where Concat(·) denotes the concatenation operation along the sequence dimension, and W_{PE} represents the learnable positional encoding. The concatenated token sequence T contains semantic information from both temporal phases while retaining the positional information of the tokens.

The structures of the Transformer encoder and decoder are shown in Fig. 3. The encoder employs a multi-head self-attention mechanism to model global dependencies between bitemporal tokens [24], as illustrated in Fig. 3(a). The core attention operation can be expressed as

$$\mathrm{Att}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where Q, K, and V denote the query, key, and value vectors, respectively, and d represents the feature dimension of each attention head. After passing through the Transformer encoder, the context-enhanced bitemporal token representation is obtained as $T_{new}=TE(T)$, where TE(·) denotes the Transformer encoder operation. The encoded token sequence is then split according to the temporal phase: $T^1_{new}, T^2_{new}=\mathrm{Split}(T_{new})$. where T^1_{new} and T^2_{new} denote the semantic tokens of the two temporal phases after global context modeling, respectively.

Subsequently, in the Transformer decoder, the pixel-level feature X^i is used as the query, while the corresponding semantic token Token T^i_{new} is used as the key and value. Through cross-attention, global semantic information is injected back into the pixel space, producing enhanced bitemporal features:

$$X^i_{new} = \mathrm{TD}(X^i, T^i_{new}), \quad i \in \{1,2\} \tag{5}$$

where $X^i_{new} \in R^{HW\times C}$ represents the enhanced feature sequence of the i-th temporal phase after decoding, and TD(·,·) denotes the Transformer decoder operation. This process enables the model to preserve global semantic modeling capability while further enhancing the discriminative representation of changed regions [25].

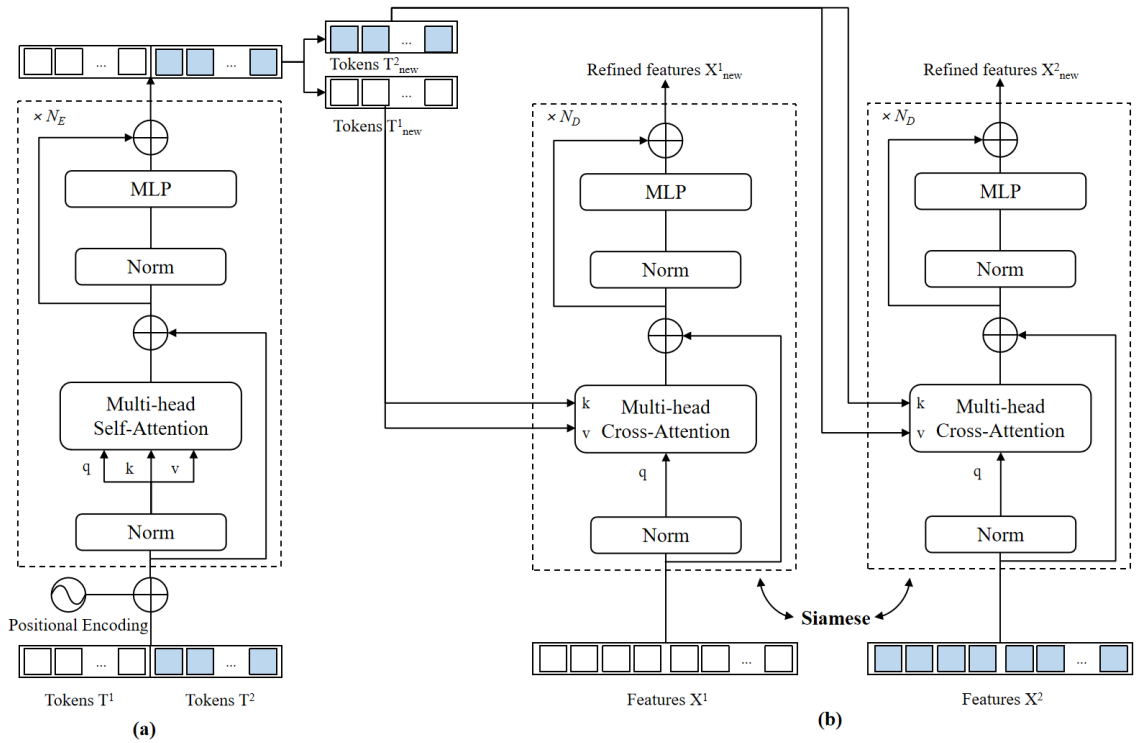


Fig. 3. Schematic illustration of the Transformer encoder and decoder used in BIT, redrawn according to the model description in Ref. [25].

B. Multi-Scale Temporal Difference Module

Although BIT [25] can effectively model the global contextual relationships between bitemporal features in the token space, its change discrimination stage mainly relies on direct differencing between the enhanced bitemporal features. Therefore, the explicit enhancement and multi-scale modeling of difference features remain insufficient. To address the challenges of high-resolution remote sensing building change detection, such as numerous small-scale targets, complex boundaries, and strong background interference, this paper introduces a multi-scale temporal difference module (TDM) after the decoder.

This module does not simply repeat the direct differencing operation in BIT; instead, it further performs multi-scale local difference enhancement based on the initial difference feature. The structure of TDM is shown in Fig. 4.

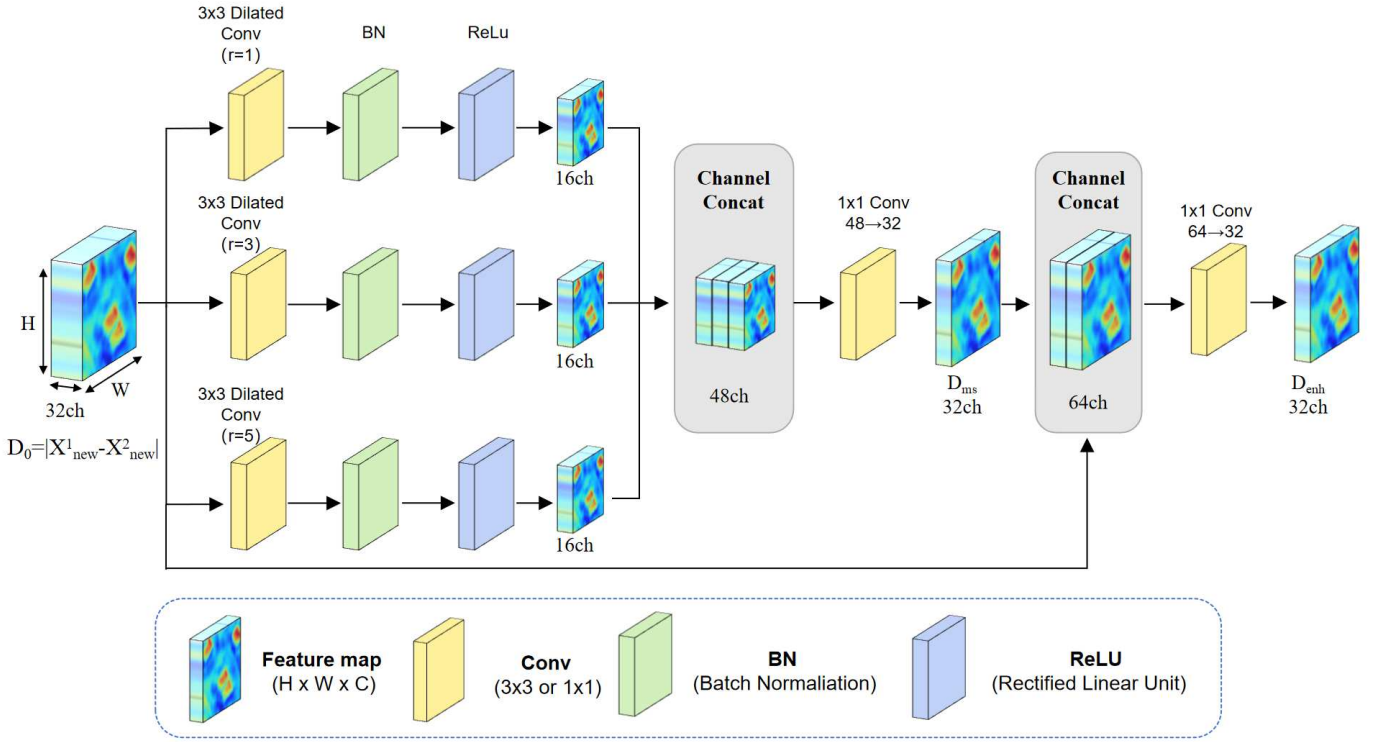


Fig. 4. Structure of the multi-scale temporal difference module (TDM).

Since TDM adopts two-dimensional convolutions for local difference modeling, the sequence features output by the decoder are first reshaped back into spatial feature maps. Let $X^i \in R^{H\times W\times C}$ denote the spatial feature map of the i-th temporal phase.

Let X^1 and X^2 denote the decoder-enhanced spatial feature maps of the two temporal images. TDM first constructs their element-wise absolute difference as the initial change representation:

$$D_0=|X^1-X^2| \tag{6}$$

Three parallel 3 × 3 convolutional branches then transform D_0 using dilation rates of 1, 3, and 5:

$$D_r = \delta\left(BN\left(Conv_{3\times3}^{(r)}(D_0)\right)\right), \quad r \in \{1,3,5\} \tag{7}$$

Here, $Conv_{3\times3}^{(r)}(\cdot)$ denotes a 3 × 3 convolution with dilation rate r, while BN(·) and δ(·) represent batch normalization and ReLU activation, respectively. The three branches have effective receptive fields of 3 × 3, 7 × 7, and 11 × 11. They therefore model local discrepancy patterns and broader structural differences without additional downsampling.

The three branch outputs are concatenated along the channel dimension and compressed by a 1 × 1 convolution:

$$D_{ms} = Conv_{1\times1}(Concat(D_1, D_3, D_5)) \tag{8}$$

where D_{ms} denotes the fused multi-scale difference feature.

Finally, the multi-scale representation is fused with the original temporal difference through a second 1 × 1 convolution:

$$D_{enh} = Conv_{1\times1}(Concat(D_0, D_{ms})) \tag{9}$$

where D_{enh} denotes the final enhanced difference representation.

The enhanced difference feature D_{enh} is subsequently passed to the prediction head, which applies upsampling and convolutional classification to produce two-channel pixel-level logits. Class probabilities are obtained by applying softmax to these logits.

C. Loss Function

Building change detection assigns either a changed or unchanged label to every pixel. However, changed buildings usually occupy only a small portion of a high-resolution image, resulting in a highly skewed label distribution. Moreover, pixels belonging to small structures, irregular contours, and transition areas are often difficult to classify correctly. To alleviate these problems, the proposed model is optimized using a joint objective that combines cross-entropy loss with focal binary cross-entropy (Focal-BCE) [54]:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{FBCE} \tag{10}$$

where λ determines the contribution of the Focal-BCE term to the overall training objective.

The cross-entropy component is formulated as

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\{0,1\}} y_{n,c}\log p_{n,c} \tag{11}$$

where N is the number of pixels, $y_{n,c}\in\{0,1\}$ is the binary indicator for the ground-truth category of pixel n, and $p_{n,c}$ denotes the probability predicted for category c. Here, c = 0 and c = 1 correspond to the unchanged and changed categories, respectively.

Although the CE term provides supervision over the complete prediction map, the optimization process may still be dominated by numerous unchanged pixels that can be recognized with high confidence. Therefore, Focal-BCE is introduced to adaptively reduce the influence of confident predictions and retain greater emphasis on uncertain pixels. It is defined as

$$\mathcal{L}_{\mathrm{FBCE}} = -\frac{1}{N}\sum_{n=1}^{N} \alpha_t\,(1-p_t)^{\gamma}\log(p_t) \qquad (12)$$
$$p_t = \begin{cases} p_n, & y_n = 1, \\ 1-p_n, & y_n = 0. \end{cases}$$

In this expression, p_n is the predicted probability of pixel n being changed, p_t denotes the probability assigned to its ground-truth label, γ is the focusing parameter, and α_t is the class-weighting factor. In our experiments, λ =0.5 γ =2.0, and α_t=1.

D. Evaluation Metrics

Model predictions are assessed at the pixel level from complementary perspectives. Precision indicates the reliability of predicted changes, Recall measures the proportion of changed pixels successfully identified, and the F1-score balances these two measures. IoU evaluates the spatial agreement between the predicted and reference change regions, while OA summarizes the correctness of predictions over both classes. Regarding the changed category as positive, TP and TN count the correctly classified changed and unchanged pixels, respectively. FP represents unchanged pixels incorrectly identified as changed, whereas FN represents changed pixels missed by the model. The five measures are computed as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \qquad (13)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad (14)$$

$$\mathrm{F1} = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \qquad (15)$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \qquad (16)$$

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \qquad (17)$$

IV. EXPERIMENTS AND RESULTS ANALYSIS

A. Datasets

Two widely used benchmarks, LEVIR-CD and WHU-CD, are selected for the experimental evaluation. Both datasets focus on building change detection using bitemporal remote sensing imagery, but they differ in image source, ground sampling distance, scene composition, and the level of structural detail available. Their complementary characteristics provide a basis for examining the robustness of the proposed method under different urban remote sensing conditions.

1) LEVIR-CD Dataset: LEVIR-CD [6] is a benchmark specifically designed for building change detection in high-resolution remote sensing images. The dataset contains 637 pairs of images acquired at two different times. Each image has a spatial size of 1024×1024 pixels and a spatial resolution of 0.5 m/pixel. The scenes mainly cover urban and suburban regions, including residential communities, industrial facilities, warehouses, and high-rise buildings.

The dataset contains considerable variations in building size and shape. In addition, buildings are often distributed sparsely, while the scenes include complex background textures and irregular object boundaries. These characteristics make LEVIR-CD suitable for assessing the ability of change detection models to identify building changes under diverse urban conditions.

For efficient training, each bitemporal image pair is partitioned into non-overlapping patches of 256×256 pixels. The original training, validation, and test partitions are retained during this process. This produces 7,120 image pairs for training, 1,024 pairs for validation, and 2,048 pairs for testing. To ensure a consistent comparison, all methods are trained and evaluated using the same data partitions. Representative examples of the original images and cropped patches are presented in Figs. 5 and Figs. 6, respectively.



(a) (b) (c)

Fig. 5. Representative original image pairs from the LEVIR-CD dataset.



(a) (b) (c)

Fig. 6. Examples of 256×256 patches generated from LEVIR-CD.

2) WHU-CD Dataset: WHU-CD [55] is constructed from publicly available aerial imagery associated with a multi-source building extraction dataset. It provides bitemporal observations of urban areas and has an approximate spatial resolution of 0.075 m/pixel. Compared with LEVIR-CD, WHU-CD offers substantially finer spatial detail, allowing building contours, narrow structures, and local texture patterns to be represented more clearly.

The higher image resolution also makes the dataset more challenging for change detection models. In particular, accurate prediction requires the model to distinguish subtle structural variations and preserve fine-grained building boundaries. This property makes WHU-CD an appropriate benchmark for evaluating detailed feature representation and boundary localization. Examples from this dataset are shown in Fig. 7.

Following the same data partitioning protocol adopted for LEVIR-CD, WHU-CD is divided into training, validation, and test subsets. These subsets contain 6,096, 762, and 762 image pairs, respectively. Since LEVIR-CD and WHU-CD differ in spatial resolution and urban scene characteristics, they provide complementary evaluation settings for building change detection. Testing the proposed method on both benchmarks enables

its performance across different high-resolution remote sensing scenarios to be examined more comprehensively.



(a) (b) (c)

Fig. 7. Representative samples from the WHU-CD dataset.

B. Experimental Settings

The model is developed using PyTorch and trained on a single NVIDIA RTX 3090 GPU with 24 GB of memory. The ResNet-18 encoder is initialized from an ImageNet-pretrained checkpoint. Model parameters are optimized by SGD with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 5×10^{-4}. The learning rate is progressively reduced according to a linear schedule. Training lasts for 150 epochs with a mini-batch size of 8. Random flipping, cropping, and blurring are applied to the training images to increase sample diversity and improve model robustness.

C. Experimental Results and Analysis

The comparison covers six established approaches with different network architectures and temporal feature interaction schemes: FC-EF [11], FC-Siam-conc [11], FC-Siam-diff [11], STANet [6], IFN [15], and BIT [25]. The first three methods mainly differ in the stage and manner in which the paired observations are combined. FC-EF merges the two images at the input level, whereas FC-Siam-conc joins their encoded features through concatenation. FC-Siam-diff instead constructs temporal representations by computing feature differences. Among the more advanced CNN architectures, STANet captures cross-temporal relationships using spatial-temporal attention, while IFN combines hierarchical features and introduces deep supervision during optimization. In contrast to these convolutional approaches, BIT employs a Transformer to model dependencies between the two acquisition times. This collection of baselines enables TDBIT to be assessed against multiple feature fusion mechanisms and network paradigms.

TABLE I.

EXPERIMENTAL RESULTS ON THE LEVIR-CD AND WHU-CD DATASETS (%)

	LEVIR-CD					**WHU-CD**				
	Pre	Rec	OA	F1	IoU	Pre	Rec	OA	F1	IoU
FC-EF	86.13	76.05	98.04	80.78	67.75	80.88	73.38	98.26	76.95	62.53
FC-Siam-conc	91.47	85.56	98.79	88.42	79.24	80.21	88.19	98.67	84.01	72.42
FC-Siam-diff	91.79	84.69	98.83	88.09	78.72	78.68	74.53	98.19	76.55	62.01
STANet	71.98	**93.13**	97.80	81.20	68.35	69.43	**92.83**	98.10	79.45	65.90
IFN	90.56	86.45	98.85	88.45	79.30	**90.22**	88.10	99.15	89.15	80.42
BIT	90.88	88.03	98.95	89.43	80.88	82.17	89.00	98.80	85.45	74.59
TDBIT	**92.08**	88.66	**98.97**	**90.34**	**82.38**	89.10	90.20	**99.17**	**89.65**	**81.24**

Table I summarizes the numerical comparison on the LEVIR-CD and WHU-CD test sets. TDBIT ranks first on both benchmarks in terms of F1-score and IoU. On LEVIR-CD, it yields an F1-score of 90.34% and an IoU of 82.38%. The corresponding values on WHU-CD are 89.65% and 81.24%, respectively. TDBIT also consistently outperforms the original BIT model. These improvements confirm the benefits of introducing explicit temporal difference enhancement and multi-scale local representation into the decoded features of BIT.

Qualitative examples from the two datasets are provided in Fig. 8 to further examine the prediction behavior of each model. The outputs show that the convolutional baselines, including FC-EF, FC-Siam-conc, and FC-Siam-diff, frequently produce false alarms or omit changed pixels near building boundaries and in small changed regions. Their predictions may also contain fragmented changed areas and incomplete contours. STANet and IFN improve the completeness of changed regions in some cases, but they still produce certain errors in complex backgrounds, dense building areas, and regions near oblique boundaries. In comparison, BIT can obtain relatively complete changed regions through global context modeling, but deviations still exist in the delineation of local boundaries and fine-grained structures.

The prediction results of TDBIT are generally closer to the ground truth, with fewer error regions. In particular, TDBIT shows more stable performance in detecting multi-scale building changes, preserving building contours, and handling complex boundary regions. These results demonstrate that further performing multi-scale temporal difference enhancement based on the initial difference of the decoded features of BIT helps enhance local difference representation and improves the detailed delineation of changed regions and the suppression of pseudo-changes to some extent.

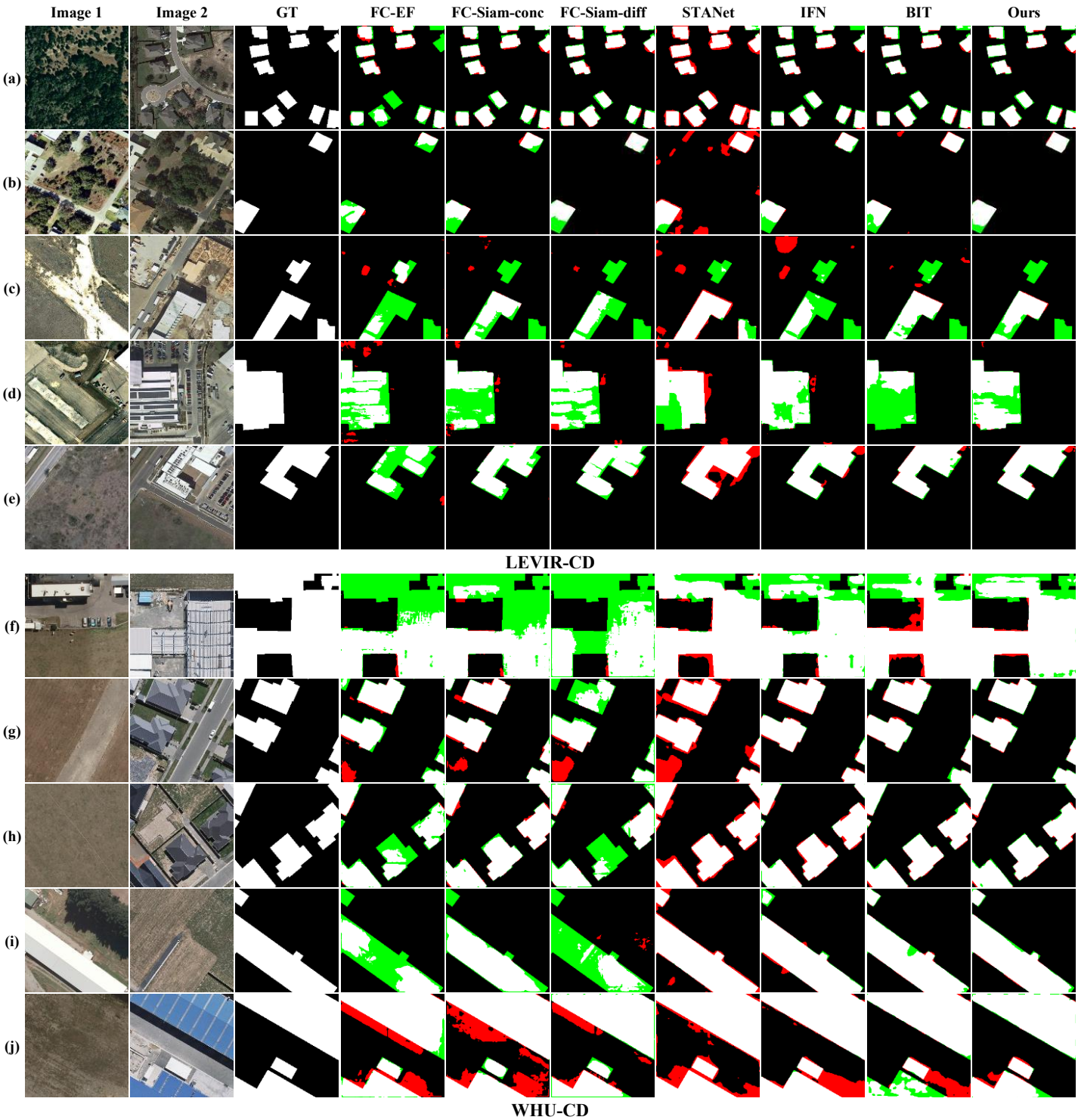


Fig. 8. Qualitative comparison on the LEVIR-CD and WHU-CD test sets. White, black, red, and green denote true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), respectively.

V. DISCUSSION

A. Ablation Study and Module Contribution Analysis

Controlled experiments on LEVIR-CD are performed to assess the roles of TDM and Focal-BCE in TDBIT. Taking BIT as the reference model, we evaluate each component independently and then examine their joint effect. The corresponding results are reported in Table II.

Augmenting the baseline cross-entropy objective with Focal-BCE increases the F1-score from 89.43% to 89.58% and the IoU from 80.88% to 81.12%. Although the gains are modest, they indicate that the modified training objective benefits change detection. By assigning lower weights to easily classified pixels, Focal-BCE directs optimization toward more challenging changed samples [54], which is consistent with its intended purpose.

When TDM is added alone, the F1-score reaches 89.81%, while the IoU rises to 81.50%. These improvements are larger

than those obtained with Focal-BCE, suggesting that TDM makes a greater contribution to the final performance. In particular, refining the initial decoder-level differences with multi-scale local information enables the network to distinguish changed regions from unchanged backgrounds more effectively.

The best performance is obtained when both components are enabled, yielding an F1-score of 90.34% and an IoU of 82.38%. Compared with the BIT baseline, these results correspond to absolute gains of 0.91 and 1.50 percentage points, respectively. The joint configuration outperforms either individual variant, indicating that TDM and Focal-BCE improve the model from complementary perspectives.

TABLE II.

ABLATION STUDY RESULTS (%).

Network	CE	Focal-BCE	TDM	Precision	Recall	F1-score	IoU
BIT	✓			90.88	88.03	89.43	80.88
BIT+Focal-BCE	✓	✓		90.49	88.68	89.58	81.12
BIT+TDM	✓		✓	90.52	89.11	89.81	81.50
TDBIT	✓	✓	✓	92.08	88.66	90.34	82.38

B. Grad-CAM Visualization and Model Response Analysis

Grad-CAM [56] is employed to identify the regions in the bitemporal images that contribute most strongly to the network's change predictions. Since change detection requires dense pixel-wise inference, the backpropagation target is defined as the average of the highest 1% of responses in the change-class logit map. The last multi-channel 2D feature layer before the final prediction layer is selected to generate the class activation maps. For a fair comparison, all heatmaps are displayed using the same normalization strategy and colormap.

From the shown visualization examples, it can be observed that, compared with BIT, the high-response regions of TDBIT are more concentrated on the changed buildings and their contour neighborhoods, while irrelevant activations in background areas are relatively reduced. This indicates that TDBIT focuses more precisely on the actual changed regions. In particular, for small-scale buildings and boundary transition regions, TDBIT produces clearer response distributions, whereas BIT exhibits more scattered responses with noticeable activations in some unchanged areas.

This phenomenon is generally consistent with the design motivation of introducing multi-scale temporal difference enhancement based on the initial difference of the decoded features and employing multi-branch dilated convolutions to model local differences at multiple scales. Qualitatively, the results suggest that the proposed method helps enhance the difference representation capability for small changed objects and complex boundary neighborhoods. It should be noted that Grad-CAM is more suitable for revealing region-level semantic responses that contribute to model decisions, rather than for precise boundary localization. Therefore, in this paper, Grad-CAM is mainly used as a qualitative visualization tool to assist in analyzing differences in attention concentration on changed regions, suppression of irrelevant background responses, and responses to small objects and boundary neighborhoods. It is not used as direct quantitative evidence of boundary modeling capability.



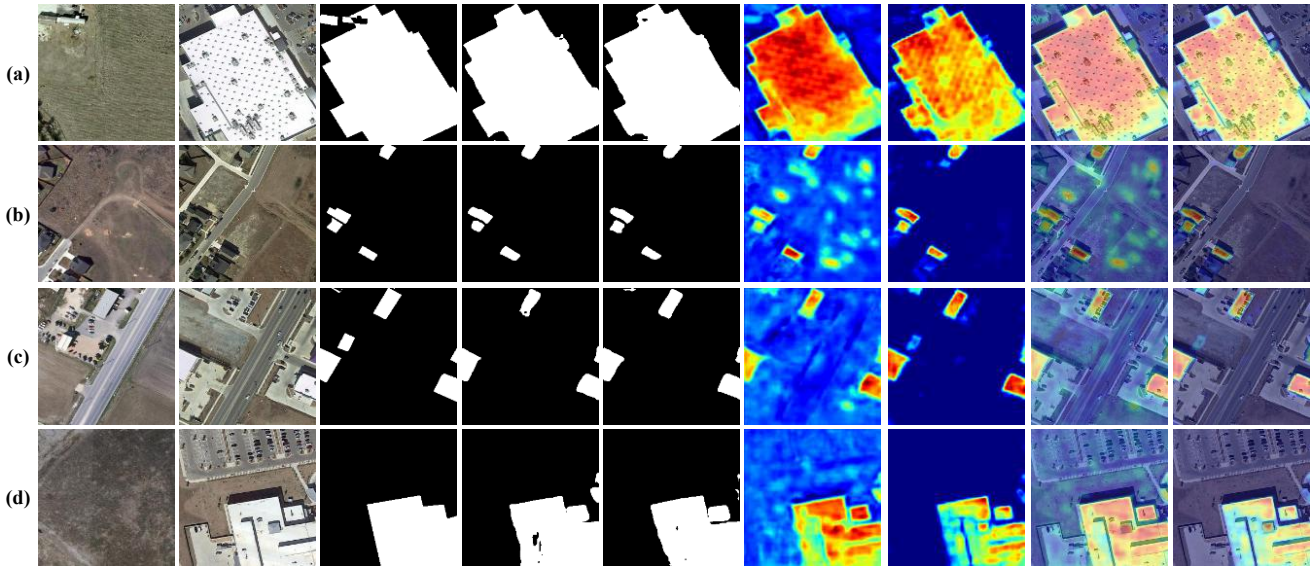


Fig. 9. Grad-CAM visualization comparison between BIT and TDBIT.

C. Limitations and Future Work

Although TDBIT performs well on test data drawn from the same distribution as the training data, its performance is still less satisfactory for very small buildings, weakly changed areas, narrow boundary structures, and objects with indistinct contours. Fine-scale components of buildings may remain undetected, suggesting that the network does not yet preserve sufficiently detailed spatial information throughout the feature extraction and reconstruction processes.

The above weaknesses are mainly associated with three factors. First, the proposed network processes image patches with a size of 256×256 pixels. The backbone produces feature maps at a lower spatial resolution than the input images. Although the output-stride adjustment and subsequent upsampling preserve more detail than the standard ResNet-18 configuration, some fine structures may still be weakened during feature extraction and prediction-map recovery.

Second, TDM is applied to the difference between the high-level contextualized feature maps produced by the decoder. Its dilated-convolution branches with different dilation rates improve the representation of changes at multiple scales, making the model more capable of emphasizing structural changes and reducing pseudo-change responses generated by complex backgrounds. However, TDM cannot completely restore small-object details that have already been weakened during the earlier stages of feature extraction. In addition, the current framework does not incorporate explicit boundary supervision, contour constraints, or a dedicated mechanism for fusing high-resolution shallow features. Consequently, the model may still produce inaccurate delineation around building edges, corners, and elongated structures. Existing research has indicated that edge-guided learning and segmentation-based change detection methods can improve boundary delineation and the completeness of changed regions in high-resolution remote sensing images [57], [58].

Third, the ability of the proposed model to generalize across substantial domain shifts requires further investigation. Variations in imaging geometry, illumination, seasonal conditions, building materials, and background texture may reduce the model's performance on unseen datasets or under more complex observation conditions [3], [5], [10]. In these situations, weak changes and small building regions are more likely to be confused with the background, thereby increasing the number of missed detections. Therefore, the cross-domain transfer capability of the current model remains limited.

In general, TDBIT is particularly effective in reducing false positives caused by complex backgrounds and strengthening the representation of multi-scale changes. However, its performance is still constrained by the inadequate recall of small objects and subtle changes, as well as its limited stability when facing data distributions that differ from those used during training. Future research will focus on incorporating high-resolution shallow-feature fusion, boundary-aware supervision, and cross-domain data augmentation or domain-adaptation techniques. These improvements are expected to enhance the robustness of fine-grained building change detection and support deployment in more complex real-world scenarios.

VI. CONCLUSION

Building changes at different spatial scales are difficult to characterize accurately in high-resolution remote sensing imagery, especially when building contours are irregular or unchanged background areas exhibit strong temporal variations. To improve the discrimination of these challenging regions, this study develops TDBIT, a building change detector derived from BIT. The proposed network retains the global contextual reasoning capability of BIT but no longer relies only on direct feature subtraction to represent temporal changes. Instead, a Temporal Difference Module (TDM) is placed after the Transformer decoder to refine the change information obtained from the two semantic feature representations. TDM first constructs an initial temporal difference map and then processes it through parallel dilated convolution branches with different receptive fields. In this way, local changes in buildings of various sizes can be captured more effectively, while structural details around irregular contours can be represented more clearly. For model optimization, the cross-entropy and Focal-BCE objectives are jointly minimized. The former supports overall pixel classification, whereas the latter assigns greater importance to

difficult samples and reduces the influence of the imbalance between changed and unchanged pixels.

The proposed method was evaluated using LEVIR-CD and WHU-CD. Compared with the original BIT network, TDBIT produced more accurate change maps on both datasets. Its F1-score and IoU were 90.34% and 82.38% on LEVIR-CD, respectively. For WHU-CD, these two metrics reached 89.65% and 81.24%. The component-wise experiments demonstrate that most of the improvement is attributable to TDM, while the joint optimization objective provides an additional benefit by strengthening the learning of difficult pixels. The Grad-CAM results also show that the responses generated by TDBIT are more concentrated within genuinely changed building areas. Consistent with this observation, the predicted maps contain fewer false alarms in unchanged regions affected by roads, shadows, exposed soil, or temporal texture variations.

Despite these improvements, TDBIT does not resolve every challenge associated with fine-grained building change analysis. Very small buildings and weak change signals may still be omitted, and precise contour recovery remains difficult in areas with narrow or irregular structures. In addition, its performance on data that differ substantially from the training distribution has not yet been sufficiently established. Subsequent research will therefore investigate the integration of shallow features with finer spatial resolution, explicit supervision of building contours, dedicated mechanisms for small targets, and augmentation or adaptation methods designed for cross-domain applications.

CREDIT Authorship contribution statement

Hao Yu: Writing – review & editing, Validation, Methodology, Conceptualization. Shenzhao Geng: Writing – original draft, Validation, Data curation. Nan Lin: Visualization, Software, Formal analysis, Resources. Yuqiang Zhang: Writing – review & editing, Supervision, Methodology. Chuanpeng Zhao: Writing – review & editing, Validation, Resources, Funding acquisition. Huiying Li: Writing – review & editing, Validation.

Acknowledgements

This work was supported by the Jilin Provincial Natural Science Foundation (Grant No. 20250101045JJ).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The LEVIR-CD dataset is publicly available at https://justchenhao.github.io/LEVIR/. The WHU-CD dataset is publicly available from the Wuhan University dataset page at https://gpcv.whu.edu.cn/data/building_dataset.html. The processed data splits and trained model weights generated in this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] A. Singh, "Digital change detection techniques using remotely sensed data," International Journal of Remote Sensing, vol. 10, no. 6, pp. 989–1003, 1989.

[2] M. Hussain, D. Chen, A. Cheng, et al., "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.

[3] Y. Wang, L. Gao, D. Hong, et al., "Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images," International Journal of Applied Earth Observation and Geoinformation, vol. 104, Art. no. 102582, 2021, doi: 10.1016/j.jag.2021.102582.

[4] N. Gonthier, "Operational change detection for geographical information: Overview and challenges," International Journal of Applied Earth Observation and Geoinformation, vol. 149, Art. no. 105278, 2026.

[5] W. Shi, M. Zhang, R. Zhang, et al., "Change detection based on artificial intelligence: State-of-the-art and challenges," Remote Sensing, vol. 12, no. 10, Art. no. 1688, 2020.

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

[7] 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.

[8] C. Wu, B. Du, L. Zhang, "Slow feature analysis for change detection in multitemporal remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 5, pp. 2858–2874, 2014.

[9] F. Thonfeld, H. Feilhauer, M. Braun, et al., "Robust change vector analysis for multi-sensor very high resolution optical satellite data," International Journal of Applied Earth Observation and Geoinformation, vol. 50, pp. 131–140, 2016.

[10] D. Peng, X. Liu, Y. Zhang, et al., "Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges," International Journal of Applied Earth Observation and Geoinformation, vol. 136, Art. no. 104282, 2025.

[11] R. C. Daudt, B. Le Saux, A. Boulch, "Fully convolutional Siamese networks for change detection," in 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4063–4067, 2018.

[12] Y. Zhan, K. Fu, M. Yan, et al., "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.

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

[14] M. Zhang, W. Shi, "A feature difference convolutional neural network-based change detection method," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7232–7246, 2020.

[15] C. Zhang, P. Yue, D. Tapete, et al., "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.

[16] S. Fang, K. Li, J. Shao, et al., "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

[17] J. Chen, Z. Yuan, J. Peng, et al., "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, 2021.

[18] H. Jiang, X. Hu, K. Li, et al., "PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," Remote Sensing, vol. 12, no. 3, Art. no. 484, 2020.

[19] H. Wei, R. Chen, C. Yu, et al., "BASNet: A boundary-aware Siamese network for accurate remote-sensing change detection," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

[20] Y. Liu, C. Pang, Z. Zhan, et al., "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, 2021.

[21] X. Wang, R. Girshick, A. Gupta, et al., "Non-local neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803, 2018.

[22] X. Peng, R. Zhong, Z. Li, et al., "Optical remote sensing image change detection based on attention mechanism and image difference," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 9, pp. 7296–7307, 2021.

[23] L. Ding, H. Guo, S. Liu, et al., "Bi-temporal semantic reasoning for the semantic change detection in high-resolution remote sensing images,"

IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022.
[24] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in Advances in Neural Information Processing Systems 30, pp. 5998–6008, 2017.
[25] H. Chen, Z. Qi, Z. Shi, "Remote sensing image change detection with transformers," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, Art. no. 5607514, 2022.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16 x 16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.
[27] S. Zheng, J. Lu, H. Zhao, et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890, 2021.
[28] N. Carion, F. Massa, G. Synnaeve, et al., "End-to-end object detection with transformers," in European Conference on Computer Vision, pp. 213–229, 2020.
[29] W. G. C. Bandara, V. M. Patel, "A transformer-based Siamese network for change detection," in 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 207–210, 2022.
[30] C. Zhang, L. Wang, S. Cheng, et al., "SwinSUNet: Pure transformer network for remote sensing image change detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.
[31] K. Zhang, X. Zhao, F. Zhang, et al., "Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.
[32] T. Yan, Z. Wan, P. Zhang, "Fully transformer network for change detection of remote sensing images," in Proceedings of the Asian Conference on Computer Vision, pp. 1691–1708, 2022.
[33] Z. Zheng, Y. Zhong, S. Tian, et al., "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 183, pp. 228–239, 2022.
[34] L. Zhang, X. Hu, Z. Mi, et al., "Object-level change detection with a dual correlation attention-guided detector," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 177, pp. 147–160, 2021.
[35] K. Yang, G. S. Xia, Z. Liu, et al., "Asymmetric Siamese networks for semantic change detection in aerial images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–18, 2022.
[36] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, 2015.
[37] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, et al., "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.
[38] L. C. Chen, Y. Zhu, G. Papandreou, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European Conference on Computer Vision, pp. 801–818, 2018.
[39] H. Zhao, J. Shi, X. Qi, et al., "Pyramid scene parsing network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890, 2017.
[40] S. Woo, J. Park, J. Y. Lee, et al., "CBAM: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision, pp. 3–19, 2018.
[41] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, 2018.
[42] Q. Shi, M. Liu, S. Li, et al., "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, 2022.
[43] Z. Zheng, Y. Zhang, Y. Wan, et al., "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 175, pp. 247–267, 2021.
[44] P. Chen, B. Zhang, D. Hong, et al., "FCCDN: Feature constraint network for VHR image change detection," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 187, pp. 101–119, 2022.
[45] H. Yang, Y. Chen, W. Wu, et al., "A lightweight Siamese neural network for building change detection using remote sensing images," Remote Sensing, vol. 15, no. 4, Art. no. 928, 2023.
[46] J. Zheng, Y. Tian, C. Yuan, et al., "MDESNet: Multitask difference-enhanced Siamese network for building change detection in high-resolution remote sensing images," Remote Sensing, vol. 14, no. 15, Art. no. 3775, 2022.
[47] J. Lei, Y. Gu, W. Xie, et al., "Boundary extraction constrained Siamese network for remote sensing image change detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.
[48] T. Liu, M. Gong, D. Lu, et al., "Building change detection for VHR remote sensing images via local-global pyramid network and cross-task transfer learning strategy," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–17, 2022.
[49] Z. Liu, Y. Lin, Y. Cao, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022, 2021.
[50] W. Wang, E. Xie, X. Li, et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578, 2021.
[51] E. Xie, W. Wang, Z. Yu, et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in Advances in Neural Information Processing Systems 34, pp. 12077–12090, 2021.
[52] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.
[53] J. Dong, W. Zhao, S. Wang, "Multiscale context aggregation network for building change detection using high resolution remote sensing images," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.
[54] T. Y. Lin, P. Goyal, R. Girshick, et al., "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988, 2017.
[55] S. Ji, S. Wei, 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.
[56] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017.
[57] A. Eftekhari, F. Samadzadegan, F. Dadrass 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, Art. no. 103180, 2023, doi: 10.1016/j.jag.2023.103180.
[58] X. Wang, X. Yan, K. Tan, et al., "Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery," International Journal of Applied Earth Observation and Geoinformation, vol. 122, Art. no. 103456, 2023, doi: 10.1016/j.jag.2023.103456.