Flow-Guided Feature Alignment and Global Attention for Fine-Grained Scene Change Detection

Yong Yang [1,2], Li C. Chen [2,*], Yun P. Liao [2], Chang Yang [1], Peng Liu [1], Li W. Zhang [3], and Qin Bao [3,*]

[1] Geological Hazard Prevention and Control Institute, Chongqing Huadi Zihuan Technology Co., Ltd., Chongqing, China
[2] Technology Innovation Center of Geohazards Automatic Monitoring, Ministry of Natural Resources, Chongqing Institute of Geology and Mineral Resources, Chongqing, China
[3] College of Computer and Information Science & College of Software, Southwest University, Chongqing, China
Email: yangyong@cqhdzh.com (Y.Y.); chenlichuandyy@163.com (L.C.C.); liaoyunping@cqdky.com (Y.P.L.); yangchang@cqdky.com (C.Y.); liupengxiao@cqdky.com (P.L.); zhang1215@email.swu.edu.cn (L.W.Z.); baoqin27@email.swu.edu.cn (Q.B.)
*Corresponding author

***Abstract*—Change detection plays a crucial role in various fields such as environmental monitoring and disaster management. In recent years, deep learning-based approaches have significantly improved detection accuracy and efficiency. However, challenges such as multi-scale feature misalignment, ineffective fusion strategies, and inconsistent cross-scale semantics still hinder their deployment in complex real-world scenarios, especially in geological hazard monitoring. To address these issues, we propose a novel change detection framework that integrates channel attention and non-local context modeling into the feature extraction stage to enhance channel discrimination and global dependency learning. In the decoding phase, we introduce a flow-guided feature alignment and fusion module, which estimates optical flow fields and performs adaptive warping to reduce temporal feature discrepancies and improve alignment accuracy. In addition, multi-level feature fusion and semantic consistency refinement are employed to better capture subtle and sparse changes. Extensive experiments on the public LEVIR-CD dataset and a newly constructed Three Gorges rock mass dataset demonstrate that our method achieves state-of-the-art performance in terms of both accuracy and computational efficiency. Moreover, the framework exhibits strong robustness in complex terrains while maintaining a lightweight design, showing great potential for practical applications in geological disaster monitoring, early warning, and risk-informed decision-making. Quantitatively, our method achieves an F1-Score of 90.0% and Intersection over Union (IoU) of 81.82% on the LEVIR-CD dataset, surpassing existing methods such as ChangeStar by 0.7% and 1.16%, respectively.**

***Keywords*—change detection, feature extraction, feature alignment, deep learning, feature fusion, attention mechanism, geological hazard monitoring**

I. Introduction

Change detection, the process of identifying variations in the state of objects or phenomena by analyzing temporally separated data, plays a fundamental role in numerous computer vision applications such as urban development monitoring, forest cover assessment, and infrastructure surveillance. In the context of geological hazard prevention, change detection is particularly critical, as it provides the technological foundation for timely early warning and risk mitigation in events such as landslides, rockfalls, and ground subsidence. Effective change detection systems can significantly reduce disaster-related losses and safeguard public safety.

With the advent of deep learning, change detection techniques have undergone rapid evolution, progressing through several developmental phases. Early works such as Fully Convolutional-Early Fusion (FC-EF) [1] introduced fully convolutional Siamese architectures for bi-temporal image comparison, while subsequent studies explored autoencoders [2], adversarial training [3], and enhanced feature reuse with U-Net++ [4]. In the architectural refinement phase (2020–2023), intermediate fusion frameworks like IFNet [5] and multi-scale networks such as LGPNet [6], SNUNet-CD [7], and a deep multi-task learning framework combining semantic segmentation with fully convolutional LSTM networks [8] improved feature representation and efficiency. More recently, attention-based approaches including MFDS-Net [9] and AM-FNet [10] have achieved notable success by incorporating global semantic enhancement and deep supervision.

Despite these advancements, existing methods face

Manuscript received April 6, 2025; revised May 28, 2025; accepted June 23, 2025; published October 24, 2025.

critical limitations in real-world scenarios, particularly under complex terrain and variable imaging conditions often encountered in geological environments. Specifically, three persistent challenges remain unresolved:

(1) Feature misalignment caused by changes in viewpoints, lighting, and temporal shifts;
(2) Cross-scale information loss during fusion of multi-level features;
(3) Insufficient global context modeling, limiting the network's ability to perceive long-range dependencies.

Recent research in visual localization and feature matching (e.g., LightGlue [11] and Unmanned Aerial Vehicle (UAV) [12]-based matching) has demonstrated the effectiveness of flow-guided alignment and context-aware feature aggregation, providing insights applicable to change detection. However, existing change detection models rarely incorporate such dynamic alignment strategies explicitly into their architectures.

To bridge this gap, we propose a novel framework that integrates flow-guided feature alignment, global attention modeling, and hierarchical fusion mechanisms. Our method is tailored for fine-grained change detection in challenging scenarios such as geological hazard monitoring, with a particular focus on robustness, accuracy, and computational efficiency.

Through extensive experiments on the LEVIR-CD [13] benchmark and a self-constructed Three Gorges rock mass dataset, we demonstrate the superior performance and practical utility of our approach in detecting subtle and sparse changes under complex environmental conditions.

II. Literature Review

Change detection has emerged as a critical task in remote sensing, with deep learning approaches revolutionizing traditional pixel-based and handcrafted feature methods. This section systematically examines recent advances in three key aspects: feature extraction architectures, feature fusion strategies, and multi-scale representation learning, while identifying persistent challenges in the field.

A. Feature Extraction Architectures

Feature extraction lies at the heart of change detection performance, evolving through several paradigm shifts. Early efforts such as FC-EF [1] employed Siamese architectures with weight-sharing encoders to capture temporal differences. Subsequent developments introduced deeper and denser models such as UNet++ [4], leveraging nested skip connections to enhance feature reuse.

Recent work has emphasized lightweight yet expressive architectures. For instance, tinyCD [14] employs depthwise separable convolutions and dual attention mechanisms to balance accuracy with computational efficiency. Attention-based models like AMFNet [10] and MFDS-Net [9] explicitly enhance long-range contextual modeling via spatial-channel attention fusion and multi-level supervision.

In parallel, recent studies in visual localization and image matching have addressed similar challenges of spatial correspondence under viewpoint and appearance variations. For instance, LightGlue achieves efficient local feature matching through dynamic graph construction and attention-based refinement, enabling accurate geometric alignment in real time [11]. Similarly, UAV visual localization frameworks evaluate the robustness of deep feature matchers in real-world aerial imagery, highlighting the importance of scale-invariant and deformation-aware representations [12]. These advances reinforce the idea that accurate temporal feature alignment—whether sparse or dense—is critical for downstream tasks. Inspired by these developments, our method integrates flow-based feature alignment to enhance spatiotemporal consistency in change detection.

Nevertheless, challenges remain in balancing the discriminative power of deep features with network complexity, particularly under severe appearance variations. This motivates the introduction of hybrid attention structures and dynamic modeling modules in our framework.

B. Feature Fusion Strategies

Effective multi-temporal feature fusion is vital for change detection. Classical paradigms include early fusion (FC-EF [1]), late fusion (SNUNet-CD [7]), and intermediate fusion (IFNet [5]). These approaches respectively focus on low-level, high-level, and multi-scale combination strategies, each with trade-offs in information preservation and model complexity.

Recent architectures such as LGPNet [6] explore hierarchical fusion with pyramid alignment to balance local and global context. Additionally, transformer-based and attention-guided methods have introduced learnable fusion gates, enhancing adaptability in diverse scenes.

Yet, two major issues persist: (1) spatial misalignment caused by temporal shifts or environmental variations; (2) static fusion mechanisms that lack context adaptivity. Inspired by the success of learnable warping and flow-based alignment in image registration and motion estimation tasks, newer models begin integrating deformable feature transformation into fusion pipelines. However, few current change detection models explicitly incorporate such flow-guided alignment for spatiotemporal consistency.

Our proposed method bridges this gap by leveraging optical flow estimation and adaptive warping in the fusion stage to mitigate inter-temporal discrepancies.

C. Multi-Scale Representation Learning

Multi-scale representation learning is essential for addressing scale variance and hierarchical semantics in change detection tasks. Early implementations, such as SNUNet-CD [7], utilized skip connections to integrate features across layers, enabling effective information flow from low-level textures to high-level semantics. MFDS-Net [9] further improved scale-awareness by introducing deep supervision at multiple levels, while tinyCD [14] demonstrated that efficient pyramidal extractors can maintain high detection accuracy even under constrained

computational resources.

To enhance scale adaptability, a variety of attention mechanisms have been proposed. For example, Squeeze-and-Excitation (SE) [15] focuses on enhancing inter-channel relationships via global pooling, while Convolutional Block Attention Module (CBAM) [16] integrates both channel and spatial attention sequentially to refine feature maps. Coordinate Attention (CA) [17] embeds location information into channel encoding, offering better spatial sensitivity, and Mixed Local Channel Attention (MLCA) [18] exploits hierarchical attention fusion across multiple scales to improve change discrimination.

While these mechanisms have achieved success in various computer vision tasks, they each present limitations when applied to high-resolution remote sensing imagery. For instance, SE and CA lack explicit modeling of spatial dependencies, while CBAM and MLCA introduce additional computational complexity that may hinder deployment in resource-constrained environments.

In this work, we adopt a lightweight Channel Attention [15] module to selectively emphasize discriminative feature channels, and further combine it with a Non-Local Block to capture long-range dependencies across the spatial domain. This hybrid design balances computational efficiency and global context modeling, which is particularly beneficial for detecting subtle or sparsely distributed changes in complex terrain. By avoiding excessive parameter overhead while maintaining expressive capacity, our attention design enables effective multi-scale feature representation and robust change localization.

III. Materials and Methods

A. Model Overview

The proposed change detection framework is designed to address the challenges of spatial misalignment, weak feature representations, and inadequate contextual modeling often encountered in multi-temporal remote sensing imagery. As illustrated in Fig. 1, the model adopts a modular architecture composed of three principal components: an encoder, a feature enhancement neck, and a decoder. Given a pair of bi-temporal optical images, the model first encodes hierarchical semantic features using a lightweight backbone network. The extracted features are then refined by the neck module, which integrates attention mechanisms to selectively enhance informative channels and model long-range dependencies. Finally, the decoder incorporates a novel Feature Displacement Alignment and Fusion (FDAF) module that explicitly addresses temporal inconsistencies through flow-guided feature alignment and fusion, ultimately producing a pixel-wise binary change map.

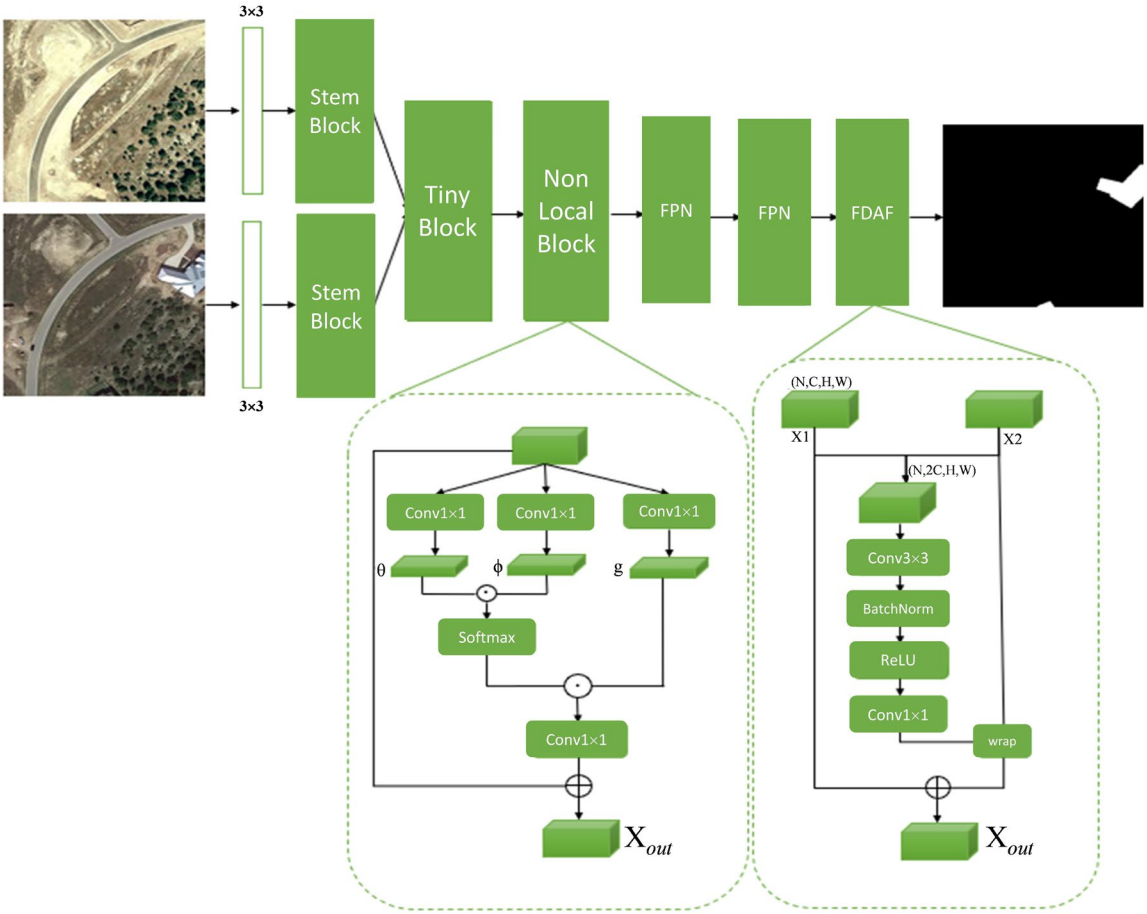

Fig. 1. The detailed architecture diagram of the model incorporates NonLocalBlock to capture non local features and strengthen global dependencies. The Feature Displacement Alignment and Fusion module (FDAF) utilizes flow fields to achieve feature displacement alignment and fusion.

The overall architecture is inspired by the encoder-decoder paradigm, but with deliberate modifications to support high-resolution inputs and preserve fine-grained details. The encoder is implemented using TinyNet [19], an efficient Convolution Neural Network (CNN) backbone that balances representational capacity with computational cost. The neck module leverages both Channel Attention and Non-Local Blocks to improve semantic discrimination and contextual awareness. The decoder employs FDAF to achieve robust change localization in the presence of object displacement, environmental variation, and geometric noise. The entire network is trained end-to-end using a cross-entropy-based objective function.

B. Attention-Enhanced Feature Extraction

The encoder is built upon TinyNet [19], which processes bi-temporal images through four hierarchical levels (stride 4 to 32 pixels) using depth-wise separable convolutions. The StemBlock module employs an inverted residual structure with 1×1 expansion (expand_ratio = 4) and 3×3 depthwise convolution, followed by PriorAttention for temporal feature fusion [19]. The TinyBlock further re-fines features through depthwise separable convolutions with optional squeeze-and-excitation attention, achieving efficient computation while maintaining feature richness [19].

The neck module contains three principal components: First, TinyFPN performs multi-scale feature fusion through lateral connections and top-down upsampling [19]. Second, the ChannelAttention module enhances feature discriminability by computing channel-wise attention weights through dual pooling paths (global average and max pooling) followed by a bottleneck MLP with reduction ratio of 16. The attention weights are generated via sigmoid activation and applied to rescale the input features. Third, the NonLocalBlock captures long-range dependencies by computing pairwise feature affinities across the entire spatial domain. It projects input features into query, key and value spaces through 1×1 convolutions (with reduction ratio of 2), computes a global similarity matrix via SoftMax normalized dot products, and aggregates contextual information through weighted summation while preserving local details via residual connections.

The decoder incorporates our novel Feature Displacement Alignment and Fusion (FDAF) module to address temporal misalignment. In the change detection task, in order to extract more discriminative features from the input image and enhance attention to key regions, we designed a feature expression module in the Neck part of the model, which improves the model's ability to capture important features through channel attention mechanism and NonLocalBlock.

The channel attention mechanism aims to enhance the expressive power of key channels by weighting the importance differences of channel features in change detection tasks. Specifically, this mechanism utilizes global average pooling and max pooling to extract two global feature descriptors, each reflecting the overall feature distribution of each channel. These two descriptors undergo a series of convolution operations and ReLU activation to generate attention weights, which are then normalized using the Sigmoid function. Finally, these weights are multiplied with the original features channel by channel to highlight the important channel information for change detection, while suppressing irrelevant or redundant features, significantly improving the model's discriminative ability for change regions. The final output features x_{out} as showed in Eq. (1).

$$x_{out} = x \cdot Attn \tag{1}$$

To further capture the global dependencies in the feature map, the model also introduces NonLocalBlocks. This module enhances the model's ability to perceive complex changes by constructing a global relationship matrix between any two points in the feature map. Specifically, the input feature map is mapped to a low dimensional space, generating three feature maps: query (θ), key (ϕ), and value (g). Query and key features are used to calculate the global similarity matrix f as showed in Eq. (2), Indicate the correlation between each location and other locations.

$$f = Softmax(\theta^T \cdot \phi) \tag{2}$$

Subsequently, the value features g is weighted and aggregated using a similarity matrix to generate global enhanced features, which are then fused with the original features through residual connections for output x_{out} as showed in Eq. (3).

$$x_{out} = W_Z \cdot (f \cdot g) + x \tag{3}$$

This non local operation not only preserves the resolution of local features, but also effectively captures long-distance global contextual information, providing powerful global modeling capabilities for change detection tasks. By combining channel attention mechanism with non local modules, the model can accurately extract local and global information from multi-scale features, thereby significantly enhancing the modeling ability and detection accuracy of changing regions.

C. Feature Displacement Alignment and Fusion (FDAF)

In change detection tasks, due to the temporal difference between the two input images, their feature distributions often experience spatial displacement. Directly fusing the two images for feature fusion may lead to accumulated errors or inaccurate detection results. To address this issue, we have designed a feature displacement alignment and fusion module in the decoder, which achieves precise alignment and fusion of feature maps at different time steps by generating flow fields. FDAF first estimates a two-dimensional flow field by analyzing feature correlations between temporal inputs through a lightweight network containing a 3×3 convolution and 1×1 convolution. This flow field is then used to warp one temporal feature set to align with the other using differentiable bilinear sampling. The aligned features are subsequently fused through element-wise addition, effectively reducing registration artifacts while highlighting genuine changes.

Specifically, the feature displacement alignment and fusion module aligns input features through optical flow estimation methods. Concatenate the two input feature maps x_1 and x_2 and input them into a convolutional network to generate a two-dimensional flow field "Flow" as showed in Eq. (4), representing the displacement information of the feature maps.

$$Flow = Conv2D(Concat(x_1, x_2)) \tag{4}$$

Subsequently, the flow field is used to distort the feature map x_2, and the features in x_2 are remapped to the coordinate system of x_1 to obtain x_2^{warp} as showed in Eq. (5).

$$x_2^{warp} = GridSample(x_2, Flow) \tag{5}$$

Finally, by fusing the aligned feature map x_2^{warp} with the original feature x_1, more robust variation features are generated x_{out} as showed in Eq. (6).

$$x_{out} = x_1 + x_2^{warp} \tag{6}$$

The feature displacement alignment and fusion module effectively alleviates the impact of differences in temporal image feature distribution on detection accuracy through flow field generation and feature distortion operations, ensuring high-quality detection of changing areas. Meanwhile, the module combines multi-scale feature maps to further enhance the model's ability to perceive subtle changes.

D. Loss Function

The loss function design of the model is based on CrossEntropy Loss, which is used to optimize the prediction results of the decoder output in the changing region. Cross entropy loss guides the model to learn more accurate category discrimination ability by measuring the difference between the predicted category probability distribution and the true labels. Its formula as showed in Eq. (7).

$$L_c = -\frac{1}{N}\sum_{C=1}^{C} y_i^c \log(\hat{y}_i^c) \tag{7}$$

Among them, y_i^c is the true label of the i-th pixel (in One Hot encoding form), and $\hat{y}_i^c$ is the predicted probability. In this task, the loss function is designed for binary classification problems (with and without changes), generating predicted probability distributions through SoftMax and comparing them with real labels to calculate errors. The use of cross entropy loss can effectively optimize the classification ability of the model, improve the detection performance of the model for changing regions by accurately measuring the difference between the predicted results and the true labels.

IV. Result and Discussion

A. Datasets

This study used two datasets to evaluate model performance. LEVIR-CD is a standard dataset for remote sensing image change detection, consisting of 637 pairs of images with a resolution of 1,024×1,024, mainly annotating the change areas of newly added and demolished buildings [13]. In addition to using the public LEVIR-CD dataset, we constructed a real-world dataset named TG-HRC, specifically designed to support change detection in geological hazard monitoring scenarios, with an image resolution of 1,920×1,080. The images were sourced from long-term monitoring video streams in the Three Gorges Reservoir area of Chongqing, China, a region known for frequent rockfall and landslide activities due to complex terrain and seasonal variation. TG-HRC covers diverse scenes including steep rock cliffs, vegetation-covered slopes, and artificial retaining structures. Image pairs were extracted under varying illumination and weather conditions to ensure temporal diversity. The change areas were manually annotated under the guidance of geological experts. The final dataset contains 12,184 training samples and 3,000 testing samples. Image registration and illumination normalization were performed as preprocessing steps. Although the dataset is representative of real-world hazardous terrain, potential biases may arise from geographic limitation (focused on a single river basin), viewpoint changes between time steps, and inherent subjectivity in manually labeling ambiguous rock shifts. To mitigate this, we applied data augmentation techniques and included varied scenes across multiple sub-regions to enhance generalization. A portion of the annotated dataset will be publicly released upon publication to support further research.

B. Evaluation Metrics

To quantitatively assess the performance of the proposed change detection model, we adopt four standard evaluation metrics widely used in binary classification tasks, particularly in remote sensing change detection: Precision, Recall, F1-Score, and Intersection over Union (IoU). These metrics jointly evaluate the model's ability to correctly identify and localize change regions, balancing the trade-off between detection completeness and reliability.

Let True Positives (TP) denote the number of correctly detected change pixels, False Positives (FP) the number of unchanged pixels incorrectly predicted as changed, and False Negatives (FN) the number of missed change pixels. The metrics are defined as follows:

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

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

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

Precision measures the proportion of detected changes that are truly changed, reflecting the model's accuracy in avoiding false alarms. Recall evaluates the model's ability to detect actual changes, indicating its sensitivity to relevant objects. F1-Score provides a harmonic mean of precision and recall, particularly useful when the dataset is imbalanced. IoU, also known as the Jaccard Index, assesses the overlap between predicted and ground-truth change areas, offering a spatial perspective of detection quality.

C. Training Strategy and Hyperparameters

Our experiment was conducted on a desktop computer running Ubuntu 20.04 operating system, which was equipped with NVIDIA TITAN V GPUs (12GB of video memory) and CUDA 10.1. Implement the use of PyTorch library [20]. During training, the model is optimized using the AdamW [21] optimizer with an initial learning rate of 0.003, momentum parameters ($\beta1$ = 0.9, $\beta2$ = 0.999), and a weight decay of 0.05. We use cosine annealing with warm restarts to enhance convergence stability.

The input image pairs are resized to 256×256, and we apply extensive data augmentation including random rotation (±15°), horizontal flipping, random cropping, brightness adjustment, contrast jittering, and time-channel swapping. The batch size is set to 16, and the training process lasts for 200 epochs.

All modules including ChannelAttention, NonLocalBlock, and FDAF are trained end-to-end from scratch. We adopt Xavier initialization for all convolutional and attention layers.

D. Quantitative Comparison

The experimental results show that the proposed model achieves optimal performance on both the publicly available dataset LEVIR-CD and the self-made Three Gorges dangerous rock dataset TG-HRC. As shown in Table I, on the LEVIR-CD dataset, the F1 value of the model reached 90.0%, and the IoU reached 81.82%. Compared with the better performing ChangeStar method, it improved by 0.7% and 1.16%, respectively, significantly improving the detection accuracy and robustness of change regions. In the TG-HRC dataset, the model also performs well in complex geological change scenarios, with an F1 value of 89.86% and an IoU of 81.59%. Both accuracy and recall are at a leading level, indicating that the model can effectively capture sparse and small change areas. The superior performance of the model is attributed to the introduced module design. The channel attention mechanism enhances the ability to capture key features, the non local module improves the perception ability of change regions by modeling global dependencies, and the feature displacement alignment and fusion module significantly reduces false alarm rates through feature alignment, providing an accurate and robust solution for change detection tasks.

TABLE I. THE QUANTITATIVE COMPARISON BETWEEN THE LEVIR-CD [11] AND TG-HRC DATASETS SHOWS THE BEST RESULTS IN BOLD. ALL RESULTS ARE DESCRIBED IN PERCENTAGE (%) FORM

Method	LEVIR-CD				TG-HRC Dataset			
	F1 (%)	Pre. (%)	Rec. (%)	IoU (%)	F1 (%)	Pre. (%)	Rec. (%)	IoU (%)
FC-EF [1]	83.4	86.91	80.17	71.53	68.38	81.2	59.05	51.95
FC-Siam-Di [1]	86.31	89.53	83.31	75.92	61.84	65.94	58.22	44.76
FC-Siam-Conc [1]	83.69	**91.99**	76.77	71.96	57.66	50.03	68.05	40.51
IFNet [5]	88.13	91.78	82.93	78.77	86.19	85.4	86.99	75.73
BIT [22]	89.31	89.24	89.37	80.68	76.79	91.19	66.31	62.32
SNUNet [7]	88.16	89.18	87.17	78.83	87.57	88.3	86.85	77.89
ChangeStar(FarSeg) [23]	89.30	89.88	88.72	80.66	75.12	88.38	65.32	60.16
Ours	**90.0**	90.05	**89.95**	**81.82**	**89.86**	**92.51**	**87.36**	**81.59**

TABLE II. COMPARISON RESULTS OF PARAMETER QUANTITY (PARAMS, M) AND COMPUTATIONAL COST (FLOPS$_s$, G). THE BEST RESULTS ARE DISPLAYED IN BOLD, WHILE THE SECONDARY RESULTS ARE SHOWN WITH AN UNDERLINE

Method	Params (M)	FLOPs (G)
IFNet [5]	50.44	82.26
SNUNet [7]	12.03	54.88
BIT [22]	<u>3.55</u>	**4.35**
Ours	**0.29**	<u>7.45</u>

The model proposed in this article not only focuses on improving accuracy in design, but also on optimizing computational efficiency and parameter scale. By introducing efficient attention mechanisms and feature alignment modules, the model achieved high accuracy metrics in change detection tasks, such as significantly improving F1-Scores and IoU metrics. However, at the same time, the FLOPs and parameter count of the model remained at a low level, reflecting the advantage of lightweight design, as shown in Table II. This design enables the model to significantly reduce the demand for computing resources while ensuring high detection performance, providing the possibility for large-scale deployment in practical scenarios.

E. Visualize Results

To provide qualitative insights into the effectiveness of the proposed method, we present visual comparisons with four representative change detection approaches—FC-EF, FC-Siam-Diff, IFNet, and SNUNet—on the LEVIR-CD dataset, as shown in Fig. 2. These results illustrate the visual differences in change localization and error patterns across various methods.

In the visualization, false positives (incorrectly

predicted change areas) are marked in red, false negatives (missed change regions) in green, and correctly detected changes in white. Compared to the baseline methods, our approach produces more precise boundaries and significantly fewer misclassified pixels, particularly in regions with small-scale or ambiguous changes. This demonstrates the benefits of our flow-guided alignment and attention-enhanced feature representation in capturing fine-grained differences across time.

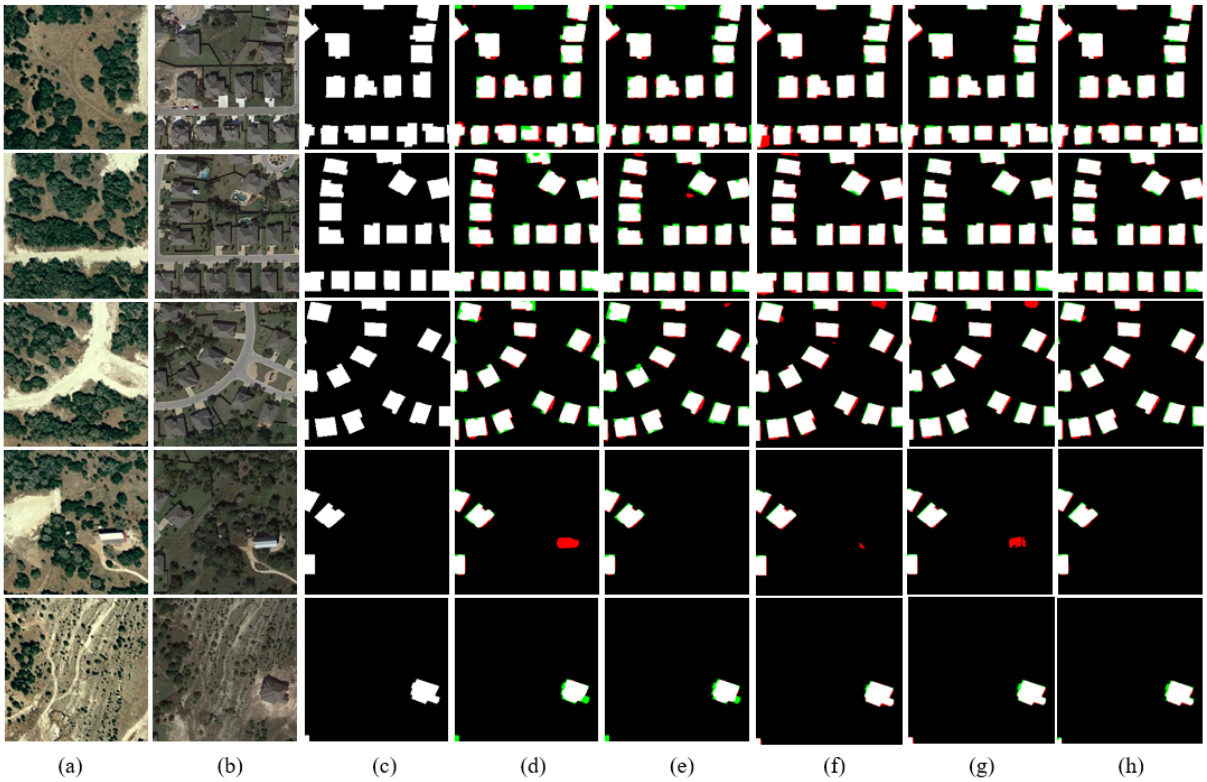

Fig. 2. Qualitative comparison of change detection results on the LEVIR-CD dataset. From left to right: (a) Image at time T1, (b) Image at time T2, (c) Ground truth, (d) FC-EF, (e) FC-Siam-Diff, (f) IFNet, (g) SNUNet, (h) Ours. False positives are shown in red, false negatives in green, and correctly detected pixels in white.

F. Ablation Experiments

We conduct extensive ablation experiments to validate the effectiveness of the proposed module: NonLocalBlock (NL) and FDAF.

All experiments are trained on the TG-HRCD dataset and evaluated on the standard validation set. The baseline model removes all three modules, while other variants incrementally add them to analyze individual contributions. Table III presents the quantitative results of different model variants, where we incrementally add each module to analyze its individual contribution.

TABLE III. ABLATION STUDY OF DIFFERENT MODEL COMPONENTS

Model Variant	F1	Pre	Rec	IoU
Baseline	89.44	92.16	86.88	80.9
Baseline+NL	89.14	91.14	87.22	80.4
Baseline+FDAF	89.57	92.33	86.97	81.11
Ours	**89.86**	**92.51**	**87.36**	**81.59**

V. CONCLUSION

This paper presents an enhanced change detection framework that incorporates Non-Local Blocks and a novel Feature Displacement Alignment and Fusion (FDAF) module to address core challenges in multi-temporal image analysis. Specifically, Non-Local Blocks strengthen global contextual modeling, improving the network's ability to capture long-range dependencies in complex terrain. The Channel Attention mechanism improves the discriminability of relevant features, while the FDAF module addresses spatial misalignment by generating flow fields and performing adaptive feature warping, thereby reducing false detections and enhancing localization accuracy.

Extensive experiments on both the publicly available LEVIR-CD dataset and our self-constructed TG-HRC geological hazard dataset demonstrate the method's superior performance in terms of precision, recall, F1-Score, and IoU, validating its effectiveness and robustness across diverse scenarios. The proposed model not only achieves state-of-the-art accuracy but also maintains a lightweight architecture suitable for real-time or embedded applications.

In terms of practical application, the method shows strong potential for integration into early warning systems for geological hazards, enabling proactive risk assessment and improved disaster preparedness.

Looking ahead, future work may focus on improving the model's generalization to multi-modal remote sensing data, such as hyperspectral imagery and LiDAR point clouds. In

addition, the incorporation of self-supervised or unsupervised learning paradigms could alleviate reliance on annotated datasets. Combining our framework with transformer-based architectures, generative adversarial training, or interpretable learning techniques may further improve its accuracy, robustness, and trustworthiness in high-risk monitoring tasks.

Overall, this work provides a solid foundation for change detection research in complex environments and opens up new directions for developing scalable, adaptable, and explainable change detection systems in the field of remote sensing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Li-Chuan Chen and Yun-Ping Liao provided guidance on research methodology, directed the research, and assisted in integrating the research content; Yong Yang, Chang Yang, and Peng Liu conducted the research, analyzed and preprocessed the data, and analyzed the research methods; Li-Wen Zhang and Qin Bao were responsible for model training, sorted the result data, and wrote the manuscript; all authors had approved the final version.

FUNDING

This work was primarily supported by the National Key Research and Development Program of China (Grant No. 2023YFC3007205). This research was also funded by the Chongqing Municipal Key Project for Technological Innovation and Application Development and the Chongqing Municipal Planning and Natural Resources Bureau Scientific Research Project, with project numbers CSTB2024TIAD-KPX0110 and KJ-2024013, respectively.

REFERENCES

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

[2] D. B. Mesquita, R. F. dos Santos, D. G. Macharet, and M. F. M. Campos, "Fully convolutional Siamese autoencoder for change detection in UAV aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1455–1459, 2019.

[3] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 45–49, 2018.

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

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

[6] T. Liu, M. Gong, D. Lu, L. Zhang, and Y. Wu, "Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.

[7] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[8] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, 2021.

[9] Z. Huang, Z. Fu, J. Song, G. Yuan, and J. Li, "MFDS-Net: Multiscale feature depth-supervised network for remote sensing change detection with global semantic and detail information," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 7507905, 2024. doi: 10.1109/LGRS.2024.3461957

[10] Z. Zhan, H. Ren, M. Xia, H. Lin, X. Wang, and X. Li, "AMFNet: Attention-guided multi-scale fusion network for bi-temporal change detection in remote sensing images," *Remote Sens.*, vol. 16, no. 10, 1765, 2024.

[11] P. Lindenberger, P. E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proc. ICCV*, 2023, pp. 17627–17638.

[12] V. V. Prutyanov, M. A. Ternov, and D. S. Kostrov, "Analysis of deep feature matching algorithms in UAV visual localization," in *Proc. 2024 Int. Russian Autom. Conf. (RusAutoCon)*, 2024, pp. 565–570. doi: 10.1109/RusAutoCon61949.2024.10694214

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

[14] A. Codegoni, G. Lombardi, and A. Ferrari, "TINYCD: A (not so) deep learning model for change detection," *Neural Comput. Applic*, vol. 35, pp. 8471–8486, 2023.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., 2018, pp. 7132–7141.

[16] S. Woo, J. Park, J. Y. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[17] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13708–13717. doi: 10.1109/CVPR46437.2021.01350

[18] D. Wan, R. Lu, S. Shen, T. Xu, X. Lang, and Z. Ren, "Mixed local channel attention for object detection," *Eng. Appl. Artif. Intell.*, vol. 123, 106442, 2023.

[19] K. Li, *et al*., "Open-CD: A comprehensive toolbox for change detection," arXiv preprint, arXiv:2407.15317, 2024.

[20] A. Paszke, *et al*., "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[21] D. P. Kingma, "Adam: A method for stochastic optimization," arXiv preprint, arXiv:1412.6980, 2014.

[22] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[23] Z. Zheng, A. Ma, L. Zhang, K. Zhong, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15193–15202.