



Article

TIMA-Net: A Lightweight Remote Sensing Image Change Detection Network Based on Temporal Interaction Enhancement and Multi-Scale Aggregation

Zhijun Zhou ¹ , Xuejie Zhang ^{1,2,*} , Xiaoliang Luo ³, Lvchun Wang ³, Wei Yu ³, Shufang Xu ¹ and Longbao Wang ^{1,2}

- ¹ College of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China; 231307050030@hhu.edu.cn (Z.Z.); xushufang@hhu.edu.cn (S.X.); wlb@hhu.edu.cn (L.W.)
² Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 211100, China
³ Jiangxi Virtual Reality Technology Co., Ltd., Nanchang 330025, China; luoxiaoliang@jx.chinamobile.com (X.L.); wanglvchun@jx.chinamobile.com (L.W.); yuwei.ict@jx.chinamobile.com (W.Y.)
 * Correspondence: xuejie_zh@hhu.edu.cn

Abstract

Remote sensing image change detection (RSCD) holds significant application value in fields such as environmental monitoring, post-disaster assessment, and urban planning. However, existing deep learning methods often face challenges of high computational complexity and insufficient detail capture, particularly demonstrating limited performance in detecting high-resolution images and complex change regions. To address these issues, this paper proposes a novel network architecture, TIMA-Net, which is designed for efficient remote sensing image change detection. By introducing the timing interaction enhancement module and the progressive decoder based on multi-scale fusion, TIMA-Net improves the accuracy of change detection while ensuring efficient computing performance. Specifically, TIMA-Net designs a temporal interaction enhancement module based on dual-branch and coordinate attention and combines channel segmentation and multi-scale features to enhance the representation ability of the changed region. The comparative experimental results show that TIMA-Net outperforms several state-of-the-art methods on multiple remote sensing datasets, especially in terms of accuracy and computational efficiency. The ablation experiment results show that each module contributes to the final performance. In summary, TIMA-Net not only provides an efficient and accurate remote sensing image change detection solution but also shows its strong potential and broad application prospects in practical applications.

Keywords: remote sensing; change detection; lightweight; deep learning



Academic Editor: Zhenwei Shi

Received: 29 May 2025

Revised: 25 June 2025

Accepted: 3 July 2025

Published: 8 July 2025

Citation: Zhou, Z.; Zhang, X.; Luo, X.; Wang, L.; Yu, W.; Xu, S.; Wang, L. TIMA-Net: A Lightweight Remote Sensing Image Change Detection Network Based on Temporal Interaction Enhancement and Multi-Scale Aggregation. *Remote Sens.* **2025**, *17*, 2332. <https://doi.org/10.3390/rs17142332>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Remote sensing image change detection (RSCD) automatically identifies surface changes by analyzing images of the same area taken at different times. This task has important application value in environmental monitoring [1], resource management [2], urban planning [3], and other fields.

Traditional change detection methods lay the foundation for RSCD tasks, including algebra-based methods and transform-based methods. Algebra-based methods usually use arithmetic operations, such as image difference [4] and image regression [5]. By calculating

the pixel difference between two-phase images, the threshold or cluster analysis [6] is used to generate the changed and unchanged difference map. Transformation-based methods, such as principal component analysis (PCA) [7] and tasseled cap transformation (KT) [8,9], map image features to specific spaces to highlight changed pixels [10]. Although traditional methods can provide effective detection results in some scenarios, they often rely on empirical feature selection and artificial threshold adjustment, and their performance is often limited in high-resolution images, especially in complex object distribution and diverse terrain scenes [11,12].

In recent years, deep learning (DL), as a powerful feature learning method, has completely changed the task of remote sensing image change detection [13]. Deep learning can achieve end-to-end change detection, and in high-resolution remote sensing images, it transcends traditional machine learning methods. Many deep learning networks (such as convolutional neural network (CNN) [14], generative adversarial network (GAN) [15], recurrent neural network (RNN) [16], transformer (Transformer) [17]) have been applied to change detection tasks. Among them, CNN has long dominated the field of change detection with its excellent feature extraction ability [18,19]. However, a major drawback of CNN is the potential loss of information, and it usually faces high computational overhead and large-scale parameters [20].

Currently, lightweight convolutional neural networks (such as MobileNetV2 [21], EfficientNet [22], etc.) have the design goals of reducing computational complexity and the number of parameters. However, they are relatively weak in feature extraction capabilities, especially in capturing detailed and semantic information. Common remote sensing image change detection methods calculate differences through direct absolute value subtraction [23] or channel-level concatenation [24]. These methods easily introduce noise and generate pseudo-changes and often fail to fully capture the complexity of temporal changes. In addition, remote sensing image change detection not only requires the accurate boundary identification of change regions but also needs to effectively fuse the spatial contextual information of images [25]. Most CNN-based methods lack long-range contextual modeling, resulting in the insufficient understanding of global semantics and background information, thus undermining the robustness of detection.

In view of the above problems, this paper proposes a lightweight remote sensing image change detection network called TIMA-Net, which aims to improve the performance of change detection in high-resolution remote sensing images. By introducing a lightweight backbone network and an innovative feature fusion strategy, the network significantly reduces the computational overhead and improves the accuracy of change detection. Compared with the existing methods, the experimental results of TIMA-Net on multiple remote sensing datasets show that it has obvious advantages in detection accuracy and computational efficiency. The main contributions of this paper are as follows:

- (1) A new lightweight change detection method, TIMA-Net, is proposed. Based on MobileNetV2 as the backbone network, Combining the Cross-Level Aggregation Fusion Module (CAFM) and the Temporal Interaction Enhancement Module (TIEM), Lightweight Multi-Scale Convolution Unit (LMCU), and Decoding Module (DM) achieves a higher performance with fewer parameters.
- (2) We proposed CAFM (Cross-Level Aggregation Fusion Module) to address the issue of disconnection between deep semantic features and shallow detail features in lightweight backbones (such as MobileNetV2), introducing a neighborhood aggregation and dynamic attention mechanism. By aggregating adjacent feature layers through sliding windows (reducing computational complexity from $O(N^2)$ to $O(N)$) and introducing adaptive weight allocation, this module alleviates the limitation of

“dimensionality explosion caused by simple concatenation” in traditional multi-scale fusion, thereby enhancing feature representation.

- (3) We proposed TIEM (Temporal Interaction Enhancement Module), which adopts a dual-branch structure to learn global and local multi-scale information of bi-temporal features respectively. It uses a joint metric of Euclidean distance and cosine similarity to replace traditional subtraction operations and introduces a Coordinate Attention (CA) mechanism to reduce noise and pseudo-changes, highlighting change regions.
- (4) We designed a lightweight progressive decoder based on multi-scale convolution, including LMCU (Lightweight Multi-Scale Convolution Unit) and DM (Decoding Module). We constructed LMCU through channel splitting and multi-directional convolutions ($1 \times 9/9 \times 1$ convolutions to capture long-range dependencies), combined with a DM module with self-attention reweighting. This decoder can extract multi-scale contextual information with fewer parameters to adapt to change features in different regions.

Compared with the existing state-of-the-art methods, TIMA-Net shows better detection accuracy and computational efficiency on three benchmark datasets.

2. Related Work

In related work, we introduced previous change detection technologies from the following three aspects: traditional RSCD methods, deep learning-based RSCD methods, and lightweight RSCD methods.

2.1. Traditional RSCD Methods

Early traditional methods for RSCD primarily focused on low- or medium-resolution images, which captured relatively simple features, such as buildings, roads, and trees, constrained by early sensor technologies. With advancements in remote sensing, various RSCD methods emerged. These techniques generated difference images (DIs) through algebraic operations, transformations, and classification, followed by thresholding or clustering to identify changed regions. For example, algebraic methods like image differencing [4], image regression [5], and clustering analysis [6] computed pixel-wise differences between bi-temporal images to produce change maps. Transformation methods, such as Principal Component Analysis (PCA) [7], perform linear transformations on multi-temporal images to extract the main components that most effectively reflect changes in the images, thereby reducing data dimensionality and enhancing the representation of change areas. The Tasseled Cap Transformation [8] is a specific image transformation method primarily used to convert spectral information from remote sensing images into indices sensitive to changes. Classification-based methods, such as Decision Trees [10], are supervised learning techniques that construct a tree-like structure by progressively dividing the image based on different features. Random Forest (RF) [12] is an ensemble learning method composed of multiple decision trees. By training several decision trees and combining their results, it further improves the accuracy of change detection.

Although these traditional methods achieved reasonable performance on low-resolution images, they heavily relied on handcrafted features and struggled with high-resolution imagery. The complex and diverse textures in high-resolution remote sensing images made it challenging for these methods to accurately capture intricate changes. Additionally, they were prone to introducing pseudo-changes or background noise under varying seasonal, illumination, and sensor conditions.

2.2. Deep Learning-Based RSCD Methods

Compared to traditional methods, deep learning approaches demonstrate superior detection performance and generalization capabilities. These methods can be categorized based on encoder–decoder architectures, such as CNN-based [26], Transformer-based [17], and hybrid CNN-Transformer models [27]. CNN-based methods, which predominantly use convolutional neural networks for feature encoding and decoding, excel in local feature extraction. For instance, three U-Net-based change detection frameworks—FC-EF [28], FC-Siam-conc, and FC-Siam-diff—explored different input configurations and fusion strategies for bi-temporal images. Subsequent research focused on enhancing change feature extraction through attention mechanisms. However, the inherent limitations of local receptive fields in CNNs hindered their ability to capture global context. To address this, deeper networks or dilated convolutions [29] were employed, albeit at the cost of increased parameters and computational complexity.

Recent efforts integrated self-attention mechanisms into CNNs to mitigate receptive field constraints. Transformers [30], with their capability to model long-range dependencies, gained traction in change detection tasks. For example, SwinSUNet [31], a pure Transformer-based U-shaped encoder–decoder, enhanced global context extraction across spatiotemporal dimensions. However, the computational overhead of self-attention remained prohibitive. Hybrid CNN-Transformer models combined CNNs' local feature extraction with Transformers' global modeling, achieving better adaptability for change detection tasks [32].

Despite their improved accuracy, deep learning methods often face challenges of high computational complexity and excessive parameters, limiting their deployment on resource-constrained platforms like UAVs or satellites. Consequently, research has increasingly focused on designing lightweight networks to balance computational efficiency and detection accuracy.

2.3. Lightweight RSCD Methods

With the complexity of remote sensing image processing tasks and the rapid growth of data volume, the computational overhead of traditional deep learning methods has become a bottleneck restricting practical applications, especially in resource-constrained environments (such as drones and embedded devices). In order to cope with this challenge, lightweight networks have gradually become an important research direction in the field of change detection, especially in application scenarios with high real-time and low energy consumption requirements.

In recent years, many lightweight networks (such as MobileNetV2 [21], ShuffleNet [33], EfficientNet [22], etc.) have been widely used in remote sensing image change detection tasks. By reducing the number of parameters and computing costs, these networks have successfully improved the efficiency of image feature extraction and significantly reduced the computational burden. For example, A2Net [34] proposed by Li et al. enhanced the expression ability of temporal features by combining progressive feature aggregation and supervised attention mechanism and achieved the efficient fusion of cross-layer features through the neighborhood aggregation module (NAM), which significantly reduced the computational overhead of the model and improved the accuracy of change detection. The USSFC-Net proposed by Lei et al. [35] optimizes the feature extraction process through multi-scale decoupled convolution (MSDConv) and the spatial–spectral feature coordination strategy (SSFC), which improves the feature representation ability while reducing the computational cost. In addition, TinyCD [36] proposed by Codegoni et al. adopts a smaller Siamese U-Net architecture, focuses on the spatial–temporal correlation of low-

level features, and improves the accuracy of change detection through spatial-semantic attention mechanisms.

Although lightweight networks perform well in change detection tasks, they still face room for improvement in performance when dealing with high-resolution remote sensing images, especially in the detection of complex change regions. SEIFNet [37] proposed by Huang et al., while maintaining lightweight design, significantly improves the ability to characterize complex change areas by introducing a spatio-temporal difference enhancement module and improves change detection accuracy and boundary integrity.

Although the existing change detection methods can maintain the accuracy of change detection while reducing the computational overhead, there is still room for further improvement in the extraction of detailed features and the reduction in the number of model parameters. Based on the existing lightweight network, our study proposes a new efficient remote sensing image change detection method by optimizing the network structure, combining temporal feature enhancement and multi-scale fusion strategy, which provides a more efficient solution for change detection tasks in resource-constrained scenarios.

3. Methods

The TIMA-Net proposed in this study uses the weight-shared MobileNetV2 to extract features and then constructs innovative modules CAFM, TIEM, LMCU, and DM. In this section, we first outline the proposed method TIMA-Net. Then, we give the specific implementation details of the proposed module. Finally, the mixed loss function is given. The architecture diagram of TIMA-Net is shown in Figure 1.

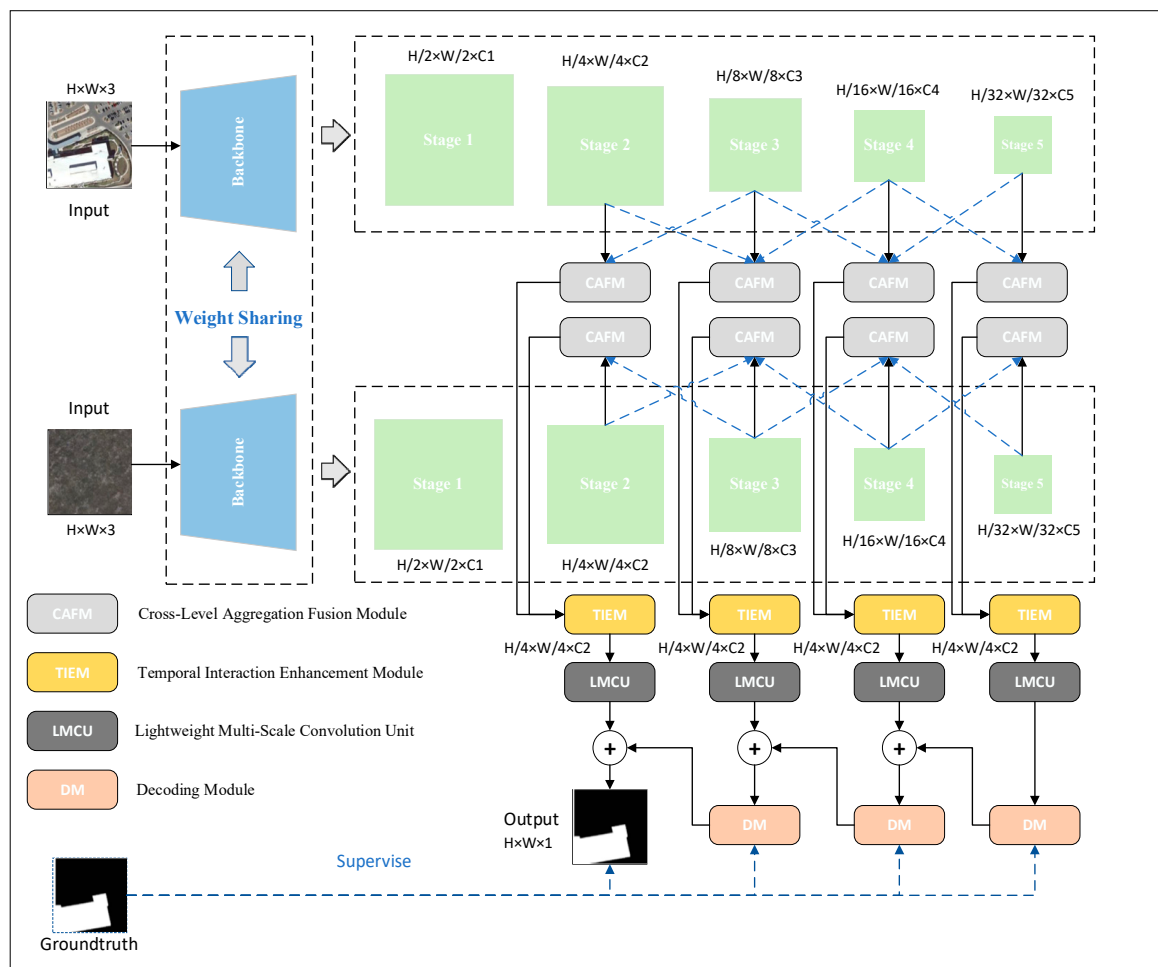


Figure 1. This figure shows the architecture of the proposed TIMA-Net.

3.1. Overall Framework

TIMA-Net is built on the popular encoder–decoder framework. The encoder includes a feature extractor, CAFM, and TIEM. The decoder includes a lightweight LMCU and a decoding module.

In the encoder, we use the lightweight backbone of MobileNetV2 as the Siamese feature extractor to extract multi-level features from low to high layers of the bi-temporal images. A pair of original images is denoted as I_A and I_B , which are captured in the same area under different conditions at time periods A and B. The corresponding pixel-wise annotated ground truth is represented by W. The hierarchical features output by the backbone for bi-temporal images can be expressed as follows:

$$F_m = \{F_m^1, F_m^2, F_m^3, F_m^4, F_m^5\}, m \in \{A, B\} \quad (1)$$

where A and B denote two periods of the same region. The backbone network extracts features of 5 levels; that is, F_m^1 represents the feature of the first level extracted from the image I_A of period A, and F_m denotes the set of features extracted by the backbone network.

To address the limitation of weak feature extraction capability in lightweight backbones, we propose the multi-scale feature fusion (CAFM) module to densely aggregate shallow detail information and deep semantic information within change regions, achieving content complementarity. This allows the network to focus more on change areas in images, particularly improving detection accuracy in complex backgrounds. The process is formulated as follows:

$$C_m^i = \text{CAFM}_i(F_m), m \in \{A, B\}, i \in \{2, 3, 4, 5\} \quad (2)$$

Among them, i indicates the feature level extracted by the backbone network, and C_m^i denotes the enhanced features at level i for the image acquired at time m .

When processing bi-temporal features, traditional methods such as absolute value subtraction, summation, and concatenation may fail to capture fine-grained changes due to ignoring temporal differences, leading to pseudo-changes affecting change detection and reducing detection accuracy. To reduce the negative impact of pseudo-changes and computational costs simultaneously, we designed a simple yet efficient Temporal Interaction Enhancement Module (TIEM) to fuse and enhance the bi-temporal features. The TIEM module extracts meaningful temporal difference information through cross-level feature comparison and the CA attention mechanism, which can significantly enhance the representation of change regions at low computational costs. The formula can be expressed as follows:

$$T_i = \text{TIEM}(C_A^i, C_B^i), i \in \{2, 3, 4, 5\} \quad (3)$$

where C_A^i and C_B^i denote the features enhanced by the CAFM module at times A and B, respectively.

In the decoder, we design a top-down decoder incorporating the Lightweight Multi-Scale Convolution Unit (LMCU) and the Decoding Module (DM) to generate the change map. By leveraging channel splitting and multi-scale convolutions in different directions, the proposed architecture enables the model to more accurately capture change information at various scales and levels of complexity in remote sensing images. The Decoding Module (DM) re-weights features through self-attention to provide precise guidance for decoding, and the formula can be expressed as follows:

$$\begin{aligned} \text{Output}_i &= \text{DMSC}(T_i) + \text{DM}(\text{Output}_{i+1}), i \in \{2, 3, 4, 5\} \\ \text{Output} &= \text{Output}_2 \end{aligned} \quad (4)$$

Finally, the output prediction and ground truth are used to compute losses in the deep supervision strategy.

3.2. Cross-Level Aggregation Fusion Module (CAFM)

In multi-scale feature representation, low-level features (e.g., F_2) excel at capturing detailed object boundaries and texture information due to their higher spatial resolution, while high-level features (e.g., F_5) retain contextual semantics critical for target localization through deep abstraction. However, conventional multi-scale fusion methods face the following three challenges: (1) simple feature concatenation causes channel dimension explosion, leading to quadratic computational complexity growth; (2) fixed-weight fusion struggles to adapt to dynamic feature contributions across diverse scenarios; and (3) independent multi-layer upsampling introduces cumulative errors, degrading detail preservation. To address these issues, we propose the CAFM, which innovates in the following three aspects: neighborhood aggregation strategy, unified resolution alignment, and dynamic attention guidance. The schematic diagram of CAFM is shown in Figure 2.

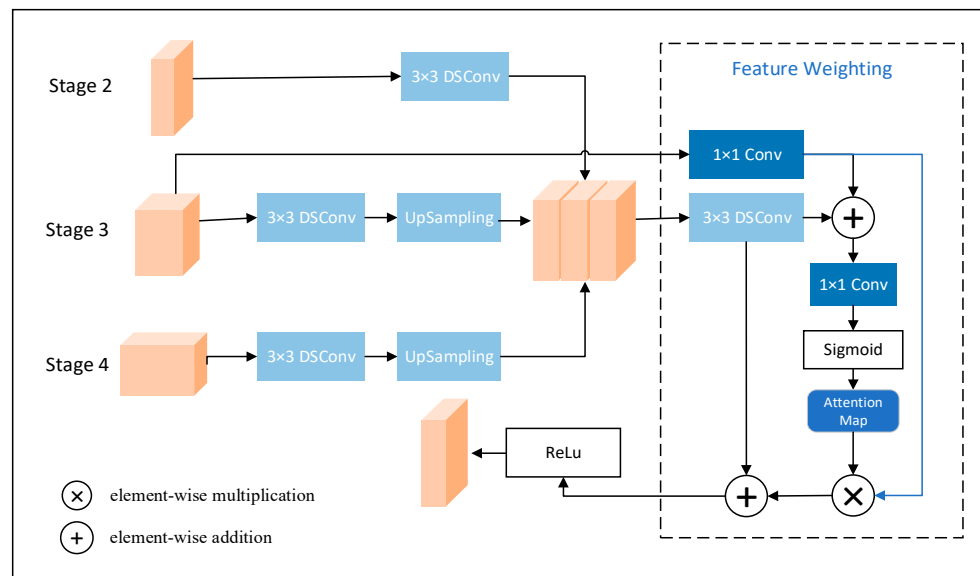


Figure 2. Illustration of the proposed CAFM.

CAFM adopts a hierarchical progressive fusion architecture, and each module only aggregates direct adjacent features to construct a local receptive field. Specifically, Submodule 1 fuses $\{F_2, F_3\}$, Submodule 2 fuses $\{F_2, F_3, F_4\}$, Submodule 3 fuses $\{F_3, F_4, F_5\}$, and Submodule 4 fuses $\{F_4, F_5\}$. This design reduces global fusion complexity from $O(N^2)$ to $O(N)$ (where N is the number of feature layers) via sliding neighborhood windows. Unlike mainstream layer-wise alignment strategies, CAFM upsamples all features to the resolution of F_2 for the following two reasons:

- (1) F_2 preserves the richest spatial details (e.g., building edges, road cracks), maximizing high-frequency information utilization;
- (2) It avoids checkerboard artifacts caused by multi-level independent interpolation.

Taking the second CAFM submodule (with F_3 as residual connection) as an example, its workflow is as follows:

$$\begin{aligned}
 f_2 &= \text{DSConv}(F_2) \\
 f_3 &= \text{UP}(\text{DSConv}(F_3)) \\
 f_4 &= \text{UP}(\text{DSConv}(F_4))
 \end{aligned} \tag{5}$$

where Up denotes upsampling, and DSConv represents a 3×3 depth-wise separable convolution with batch normalization (BN) and ReLU activation.

We then concatenate the features f_2, f_3 , and f_4 along the channel dimension to enrich the feature representation. We use a 1×1 convolution for residual learning to retain the key information of f_3 and adjust the channel number of the concatenated features using a 3×3 depth-wise separable convolution. This process can be expressed as follows:

$$\begin{aligned} F_i &= \text{Conv}1 \times 1(f_3) \\ F_c &= \text{DSConv}3 \times 3(\text{Concat}(f_2, f_3, f_4)) \end{aligned} \quad (6)$$

Among them, Concat represents the feature concatenation operation, and Conv 1×1 represents a 1×1 convolutional layer with batch normalization (BN) and ReLU activation function.

Finally, a simple dynamic weight mechanism is introduced. After performing element-wise addition on F_i and F_c , an attention map is generated through a 1×1 convolution and a Sigmoid activation function. Then, this attention map is used to weight and fuse the feature F_i ; that is, the importance of features at each position is adjusted according to the output of the attention mechanism. If the value of the Attention_map at a certain position is large (close to 1), the feature at that position will be assigned a higher weight; otherwise, it will be weakened. Then, the weighted original feature $F_i \otimes \text{Attention_map}$ is added element-wise to the fused feature F_c to generate a new feature map F_{output} . The main operations can be expressed as follows:

$$\begin{aligned} F_{\text{fused}} &= F_i \oplus F_c \\ \text{Attention_map} &= \text{Sigmoid}(\text{Conv}1 \times 1(F_{\text{fused}})) \\ F_{\text{output}} &= \text{ReLU}(F_i \otimes \text{Attention_map} \oplus F_c) \end{aligned} \quad (7)$$

where $\oplus$ represents element-wise addition, and $\otimes$ denotes an elementwise multiplication operation. Conv 1×1 represents a 1×1 convolutional layer with batch normalization (BN) and ReLU activation function.

Each CAFM adopts a similar operation. The only difference lies in the features used for the residual connection. We obtain four levels of bi-temporal enhanced features. This design weights the original features and fused features through an attention mechanism to enhance the expression of key features, while retaining original information via residual connections. This approach improves the model's feature representation capability and robustness.

3.3. Temporal Interaction Enhancement Module (TIEM)

Currently, many methods focus solely on the extraction of multi-level features but often overlook the interaction between bi-temporal features within the same level. Simple operations like direct subtraction or summation are commonly used to fuse the bi-temporal features. However, although subtraction-based methods can reduce computational complexity, they often lead to the generation of pseudo-changes. The concatenation-based methods retain the features in the channel dimension, but they generate excessive redundant information, making it difficult for the model to pay more attention to the changed areas.

To address this issue, we designed the TIEM. The TIEM learns both differential features and concatenated features simultaneously, enabling it to obtain global change related information and fine-grained local context. The schematic diagram of TIEM is shown in Figure 3.

We first perform element-wise addition on features in the channel dimension and utilize two convolutional blocks to extract local contextual information for detail supplementation and noise interference suppression. The difference branch has innovatively

improved the traditional direct subtraction operation; the difference features are calculated by fusing the Euclidean distance and cosine similarity. Remote sensing images usually have the characteristics of wide coverage and are susceptible to factors such as illumination, seasons, and weather. The changes of ground objects in change detection tasks may manifest as abrupt amplitude changes, morphological evolutions, or spatial displacements, and these changes are often accompanied by complex background noise interference. Feature fusion methods, such as traditional simple subtraction, are difficult to effectively capture such complex change features and are prone to generating pseudo-change regions.

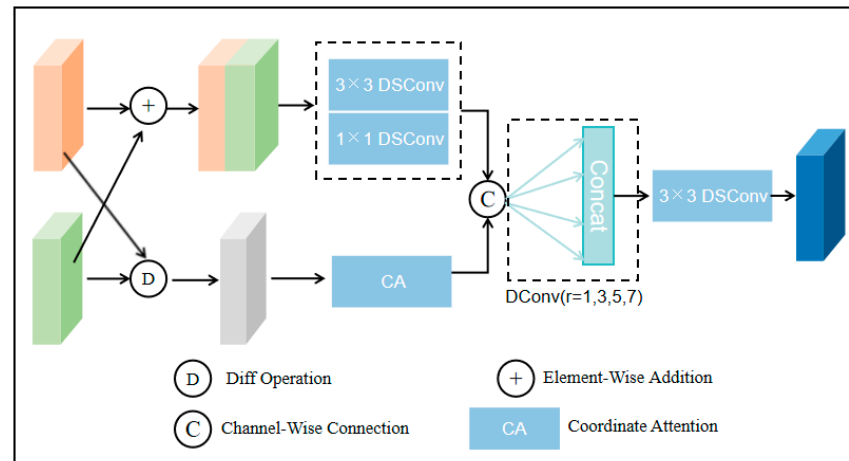


Figure 3. Illustration of the proposed TIEM.

From the perspective of theory and mathematical principles, the combination of Euclidean distance and cosine similarity has significant advantages. First, it is the joint representation of amplitude and direction; the Euclidean distance ($D_{\text{euc}} = \sqrt{\sum_{i=1}^N (X_1^i - X_2^i)^2}$) quantifies the geometric distance of feature vectors, which is sensitive to changes, such as abrupt spectral intensity changes, caused by building demolition; the cosine similarity ($S_{\text{cos}} = \frac{X_1 \cdot X_2}{\|X_1\| \|X_2\|}$) measures the directional consistency of vectors, which can resist the same-direction amplitude noise caused by illumination fluctuations of ground objects. Combining the two can capture both “amplitude differences” and “directional differences”. When water pollution causes the spectral curve to tilt (the amplitude remains unchanged but the direction changes) or the building shadow changes (the direction remains unchanged but the amplitude changes), this combination can more accurately characterize real changes, and the F1 score is increased by 0.63% compared with single subtraction.

We also adopt a new attention mechanism proposed by Hou et al., called Coordinate Attention (CA) [38]. This mechanism embeds positions into channel attention, which can balance efficiency and spatial location perception with a relatively low number of parameters.

The expressions for the concatenated feature Con and the difference feature Diff. Here, the difference feature Diff is obtained by concatenating the Euclidean distance and cosine similarity along the channel dimension, which can be expressed as follows:

$$\begin{aligned} \text{Con} &= \text{DSConv}1 \times 1(\text{DSConv}3 \times 3(X_1 \oplus X_2)) \\ D_{\text{euc}} &= \sqrt{\sum_{i=1}^N (X_1^i - X_2^i)^2} \\ S_{\text{cos}} &= \frac{X_1 \cdot X_2}{\|X_1\| \|X_2\|} \\ \text{Diff} &= \text{CA}(\text{Cat}(D_{\text{euc}}, S_{\text{cos}})) \end{aligned} \quad (8)$$

Among them, $(X_1 \oplus X_2)$ represents the element-wise addition operation, $X_1 \cdot X_2$ represents the dot product of X_1 and X_2 , $\|X_1\| \|X_2\|$ represents the norm of X_1 and X_2 , CA represents the coordinate attention module, and Cat represents concatenation along the channel dimension. DSConv 3×3 and DSConv 1×1 represent 3×3 depth-separable convolution operations and 1×1 depth-separable convolution operations with batch normalization (BN) and ReLU activation functions, respectively.

The calculation method combining Euclidean distance and cosine similarity has significant advantages in feature difference measurements. The Euclidean distance can effectively capture the geometric distance between features and retain the information of spatial position changes between features, enabling the model to identify the details of amplitude changes. The cosine similarity, on the other hand, focuses more on the directional differences between features and can reflect the relative directional changes of features, even when the amplitude differences between them are small. By combining these two measurement methods, the model can simultaneously focus on both amplitude and direction differences, thereby enhancing its ability to perceive details and global changes.

In addition, after concatenating the concatenated features and differential features in the channel dimension, the TIEM sets up four parallel branches, each with a different dilation rate $r \in \{1, 3, 5, 7\}$, to expand the receptive field and capture changes at different scales. This design of multi-scale perception enables the model to learn change features from different scales, thus better identifying changes of various scales.

In summary, the TIEM module significantly improves the ability to capture change features through its innovative dual-branch structure, the difference calculation method combining Euclidean distance and cosine similarity, and the introduction of the coordinate attention mechanism. This module can not only effectively extract global change information but also further enhance the model's performance through fine-grained local context learning and multi-scale perception strategies. The TIEM demonstrates strong advantages in change detection tasks, with efficient computational performance and accurate change recognition ability.

3.4. Lightweight Multi-Scale Convolution Unit (LMCU)

In the context of change detection tasks, existing methods often encounter several challenges. Firstly, traditional models typically rely on single-scale feature representations, which limits their ability to effectively capture multi-scale contextual information within complex scenes. For instance, when processing high-resolution satellite imagery, focusing solely on local details may overlook crucial global structural information. Secondly, many current solutions tend to employ large convolution kernels or stack multiple convolutional layers to enhance the receptive field. However, this inevitably increases computational overhead and model complexity, leading to prolonged training times and higher resource consumption.

To address these issues, this paper proposes a Lightweight Multi-Scale Convolution Unit (LMCU), designed to enhance the model's understanding of multi-scale contextual information while maintaining computational efficiency. By introducing the combination of depth-wise separable convolution and convolution kernels of different sizes, LMCU can effectively capture multi-level context information from local to global, thus enhancing the accuracy and fine-grained perception of change detection. Compared with traditional methods, LMCU not only significantly improves the recognition ability of the model to complex scene changes but also reduces the computational overhead through efficient design, making the model more suitable for real-time applications in resource-constrained environments. The schematic diagram of LMCU is shown in Figure 4.

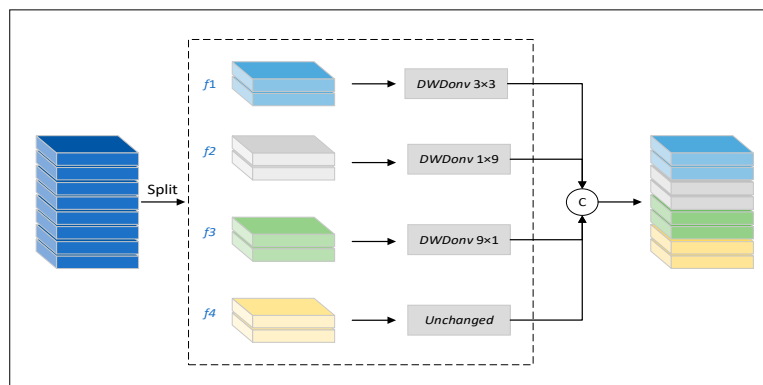


Figure 4. Illustration of the proposed LMCU.

Firstly, the input feature map is split into multiple sub-feature maps along the channel dimension (as shown in the figure, f_1 , f_2 , f_3 , and f_4). The first branch processes sub-feature map f_1 using a 3×3 depth-wise separable convolution kernel, which facilitates the capture of detailed features within local regions. The second branch f_2 employs a 1×9 convolution kernel to process sub-feature map, focusing on extracting long-range dependencies in the horizontal direction. The third branch utilizes a 9×1 convolution kernel to handle sub-feature map f_3 , aiming to identify global associations in the vertical direction. The fourth branch directly passes sub-feature map f_4 without modification, preserving the original feature information and ensuring that critical details are not lost.

After processing by these four branches, the resulting feature maps are concatenated along the channel dimension for integration, generating the final output feature map. By incorporating convolution kernels of different sizes, the LMCU effectively captures multi-scale contextual information ranging from local to global scales, thereby enhancing the model's ability to perceive changes in complex scenes. Despite the introduction of multiple convolutional branches, the design of channel splitting and depth-wise separable convolutions allows the LMCU to achieve performance improvements, while maintaining efficient computation.

This design ensures that the LMCU can efficiently extract rich feature representations at various scales without significantly increasing computational complexity, making it particularly suitable for tasks requiring high accuracy and efficiency in change detection.

3.5. Decoding Module (DM)

In change detection tasks, accurately extracting and enhancing change features is the key to achieving efficient detection. To overcome the noise and inconsistency issues faced by traditional methods in multi-scale feature fusion, we propose an innovative Decoding Module (DM). This module combines the self-attention (SA) and the inverse operation method to effectively weight the change regions, thereby improving detection results. The schematic diagram of DM is shown in Figure 5.

In the DM, the model first utilizes the SA to dynamically focus on relevant features, capturing long-range dependencies and enhancing the representation of important areas in the image. The core of this process lies in generating Query, Key, and Value through convolution operations and calculating the similarity between them to obtain the attention weight of each position. The feature map d_3 output from the SA is processed by a 1×1 convolution layer and mapped to the range of $[0, 1]$ through the Sigmoid activation function to generate the change map c_3 . Assuming the input feature of this module is d_{input} , the process is expressed as follows:

$$c_3 = \text{Sigmoid}(\text{Conv1} \times 1(\text{SA}(d_{input}))) \quad (9)$$

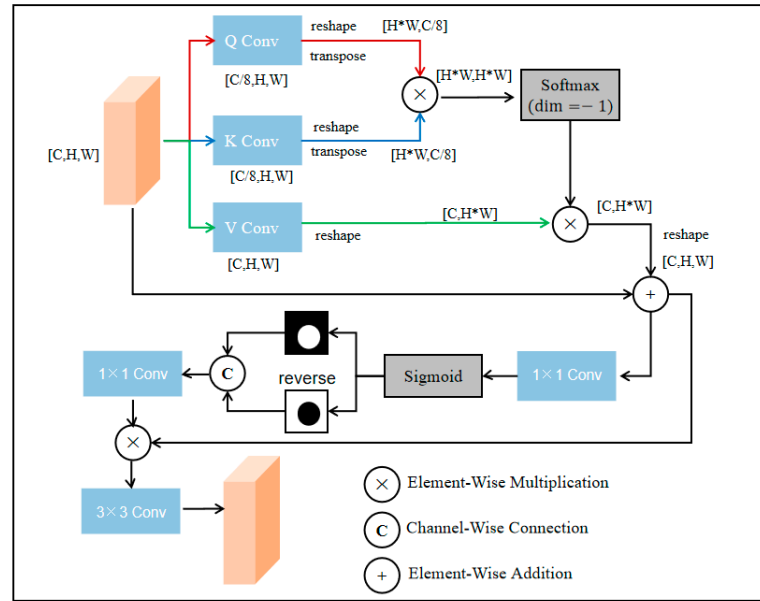


Figure 5. Illustration of the proposed DM.

To further enhance the ability to distinguish between the background and change regions, we introduce an inverse operation module. The inverse operation generates the inverse change map $c3'$ by reversing the change map $c3$. This process is accomplished through a subtraction operation. The generated inverse change map $c3'$ has opposite weights to the change map $c3$, which can effectively eliminate background information and retain only the features of the change regions, as follows:

$$c3' = \text{reverse}(c3) \quad (10)$$

After the inverse operation, we concatenate the change map $c3$ and the inverse change map $c3'$ and then further process them through a 1×1 convolution to generate a comprehensive change mask $a3$. Finally, we use the generated attention mask $a3$ to weight the feature map $d3$. The feature map is refined and rescaled by element-wise multiplication, and finally, a 3×3 convolution is used again for feature extraction and refinement.

$$\overline{d3} = \text{Conv}3 \times 3 (\text{Conv}1 \times 1 (\text{Cat}(c3, c3')) \otimes d3) \quad (11)$$

This operation can effectively enhance the perception of the change regions and suppress background noise. Through this mechanism, the supervised attention module can clearly distinguish the change information from the background information, thereby enhancing the robustness and accuracy of change detection.

3.6. Loss Function

For change detection tasks, in most cases, the proportion of change regions is much smaller than that of unchanged regions, which leads to the problem of class imbalance. To mitigate this issue and guide the network to learn from complex scenarios, we adopt a hybrid loss, including the Binary Cross Entropy (BCE) loss L_{BCE} and the Dice loss L_{Dice} [39]. The BCE loss can be expressed as follows:

$$\mathcal{L}_{bce}(p, \bar{q}) = \bar{q} \cdot \log p + (1 - \bar{q}) \log(1 - p) \quad (12)$$

In the formula, $\cdot$ represents the dot product operation, and $p, \bar{q}$ are the predicted change map and the corresponding ground truth, respectively. The Dice loss can be expressed as follows:

$$\mathcal{L}_{Dice}(p, \bar{q}) = 1 - \frac{2 \cdot p \cdot \bar{q}}{\|p\| + \|\bar{q}\|} \quad (13)$$

where $\|\cdot\|$ represents the L1 norm. The total loss is expressed as follows:

$$\mathcal{L}_{Total}(p, \bar{q}) = \sum_{i=2}^5 \{\mathcal{L}_{bce}(p_i, \bar{q}) + \mathcal{L}_{Dice}(p_i, \bar{q})\} \quad (14)$$

Among them, $i \in \{2, 3, 4, 5\}$, and they, respectively, represent the predicted change maps at four stages.

4. Results

4.1. Configuration

4.1.1. Implementation Details

We use the PyTorch 1.8.0 framework to implement the proposed method and use a single 24 GB NVIDIA RTX 3090 GPU for training/inference.

In the training parameter settings, we employed the Adam optimizer with a weight decay of 1×10^{-4} , an eps parameter of 1×10^{-8} , and momentum parameters of 0.9 and 0.99, respectively. The initial learning rate was set to 5×10^{-4} , the batch size was 16, and the max_iteration was configured as 40,000. To further enhance the model's generalization capability, several data augmentation techniques were adopted during the training phase, including normalization, random scaling, cropping, flipping, and exchanging. All images were uniformly scaled to a resolution of 256×256 , randomly cropped into small patches of approximately 8×8 pixels, and flipped/exchanged randomly with a 50% probability.

4.1.2. Datasets

We use three benchmark datasets for remote sensing image change detection, as follows:

LEVIR [40]: This is a building change detection dataset, consisting of 637 pairs of bi-temporal RS images with a spatial size of 1024×1024 and a spatial resolution of 0.5 m. We crop the images into 256×256 patches with overlaps and split the dataset in a 7:1:2 ratio, resulting in 7120/1024/2048 image pairs for training/validation/testing.

SYSU [41]: This dataset contains 20,000 pairs of bi-temporal RS images with a spatial size of 256×256 and a spatial resolution of 0.5 m. We split the dataset in a 6:2:2 ratio. This dataset encompasses a variety of complex change scenarios, such as road widening, the development of new urban buildings, changes in vegetation, and suburban growth.

BCDD (http://study.rsgis.whu.edu.cn/pages/download/building_dataset.html, accessed on 7 November 2024): This building change detection dataset contains a pair of images with a size of $32,507 \times 15,354$ and a high resolution of 0.075 m. We cropped the images and overlapped them into patches of size 256×256 . Then, we obtained 6095/762/763 pairs of images for training/validation/testing.

4.1.3. Evaluation Metrics

We adopt four widely used evaluation metrics, namely intersection over union (IoU), F1-score (F1), recall (Rec), and precision (Pre), to evaluate the performance of remote sensing image change detection. Among them, Tp, Fp, Tn, and Fn represent the numbers of true positives, false positives, true negatives, and false negatives, respectively.

4.2. Comparative Studies of SOTA Methods

We compare the proposed model with seven state-of-the-art RSCD methods, including five CNN-based methods: FC-Siam-diff (ICIP, 2018) [28], DMINet (TGRS, 2023) [42] and SEIFNet (TGRS, 2024) [37], and Transformer-based method: BIT (TGRS, 2021) [17]. In addition, the methods ICIF-Net (TGRS, 2022) [30], STFF-GA (TGRS, 2024) [43], and RaHFF-Net (JStars, 2025) [44] combining CNN and Transformer are also considered.

The model parameters and computational cost of the above method are also reported. All experiments were conducted on NVIDIA GeForce RTX 3090 (NVIDIA Corporation, Santa Clara, CA, USA).

4.2.1. Quantitative Comparison

Table 1 reports the quantitative comparison of IoU, F1, Rec, and Pre of different methods on two RSCD datasets. We used F1 and IoU as the main evaluation indicators. The top three values are marked in bold red, bold green, and bold blue, respectively. We also reported the size of model parameters (Params) and computational costs (FLOPs) in Table 1.

Table 1. Quantitative comparisons of IoU, F1, Rec, and Pre on three RSCD datasets.

Methods	LEVIR (%)				SYSU (%)				BCDD (&)				FLOPs (G)	Params (M)
	IoU	F1	Rec	Pre	IoU	F1	Rec	Pre	IoU	F1	Rec	Pre		
FC-Siam-diff	75.18	85.84	84.59	87.12	52.06	68.47	55.89	88.36	71.94	83.68	77.76	90.57	9.42	1.35
BIT	82.76	90.57	89.40	91.77	65.84	79.40	76.68	82.32	88.38	93.83	92.03	95.70	16.88	3.01
ICIF-Net	82.03	90.13	88.89	91.40	62.20	76.70	76.99	76.39	88.45	93.88	92.12	95.61	25.41	23.84
DMINet	82.92	90.67	88.89	92.54	68.11	81.19	78.10	84.19	88.72	93.95	92.30	95.65	14.55	6.24
STFF-GA	83.39	90.93	91.34	90.55	69.45	81.97	80.14	83.89	89.11	94.24	92.68	95.82	8.22	12.28
SEIFNet	83.40	90.95	89.46	92.49	69.96	82.32	79.98	84.81	88.95	94.10	92.55	95.75	8.37	27.91
RaHFF	80.85	89.41	89.35	89.46	66.73	80.04	79.79	80.30	81.86	90.03	86.64	93.68	33.95	16.48
AFRNet (ours)	83.96	91.28	89.97	92.62	70.49	83.01	80.31	85.88	89.43	94.42	93.65	95.84	6.05	4.31

The method we proposed achieved competitive scores on all four evaluation metrics on the three datasets. A more detailed description is that the IoU/F1/Rec/Pre scores of TIMA-Net on the LEVIR-CD dataset are 83.96%/91.28%/89.97%/92.62%. It exceeds 80.85%/89.41%/89.35%/89.46% of the recently proposed RaHFF and is superior to 83.40%/90.95%/89.46%/92.49% of SEIFNet.

On the LEVIR-CD dataset, the RaHFF method proposed in 2025 requires 16.48 M parameters and 33.95 G FLOPs to obtain 89.41% of the F1 measurement, while the method proposed in this paper only requires 4.31 M parameters and 6.05 G FLOPs to obtain 91.28% of the F1 measurement. The main reason lies in that the CAFM module innovatively enhances the fusion effect of low-level and high-level features while maintaining computational efficiency through the neighborhood aggregation strategy and dynamic attention guidance. This design significantly enhances the feature representation ability of TIMA-Net in complex backgrounds, avoiding the explosion of computational complexity and information loss caused by simple feature concatenation in the RaHFF method, thereby improving the accuracy of change detection. Compared with the lightweight method SEIFNet, the method proposed in this paper improves the detection accuracy while reducing the number of parameters. The main reason lies in that the method proposed in this paper further enhances the detail perception of the changed area through the self-attention mechanism in the DM module and successfully optimizes the detail representation of the change graph. In the detection of change areas, TIMA-Net can better distinguish the background from the change areas, effectively suppress noise, and significantly improve the accuracy of change detection.

On the SYSU dataset, the IoU, F1, Rec, and Pre of the method proposed in this paper are always superior to all other methods. Compared with the STFF-GA proposed in 2024, although the method in this paper only slightly leads the STFF-GA in detection accuracy, the number of parameters in this paper is reduced to one-third of that of the STFF-GA method. The main point is that the STFF-GA method uses complex convolution operations and feature fusion strategies, but this design leads to a relatively high number of parameters and computational overhead. In contrast, the method proposed in this paper pays more attention to lightweight in design. Especially through the CAFM and TIEM modules, unnecessary computations have been successfully reduced, and excellent detection performance has been maintained through efficient multi-level feature fusion. Compared with other methods, the method proposed in this paper also achieves the best performance with almost the lowest model parameters and cost. All these indicators strongly demonstrate the superiority of the proposed lightweight TIMA-Net method.

4.2.2. Qualitative Evaluation

We show the visual comparisons between the proposed TIMA-Net and different sota methods on three RSCD datasets in Figures 6–8. We use several colors to represent true positives (white), false positives (red), true negatives (black), and false negatives (blue).

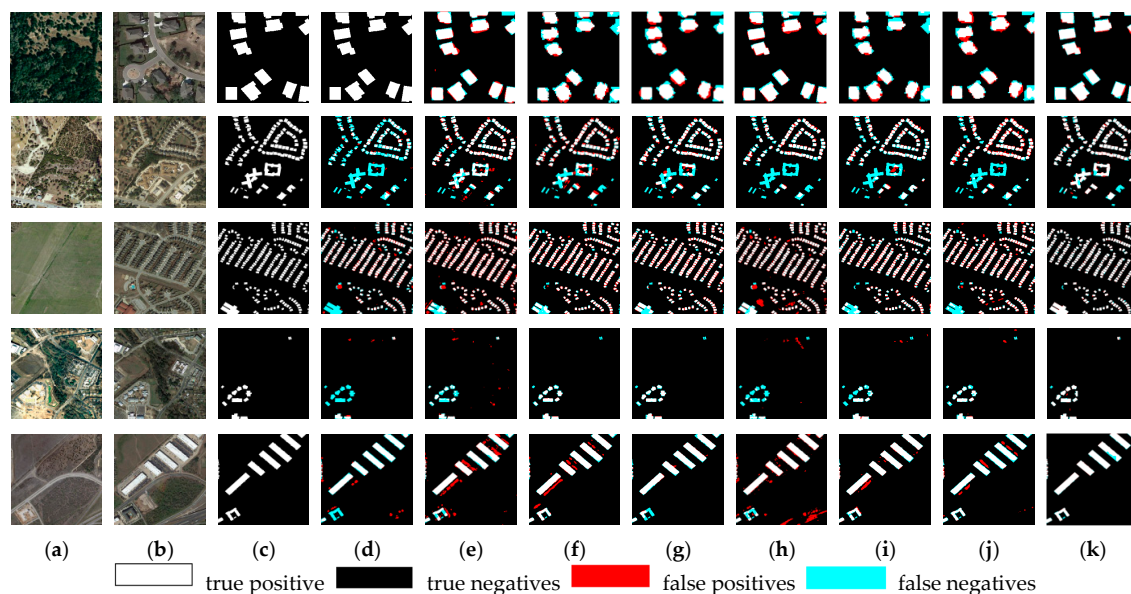


Figure 6. LEVIR dataset. (a) t1 images (b) t2 images (c) Ground truth (d) FC-Diff (e) BIT (f) ICIF-Net (g) DMINet (h) STFF-GA (i) SEIFNetNet (j) RaHFF (k) Ours.

The results indicate that the method proposed in this paper exhibits superiority in the following aspects:

- (1) Advantage in reducing the occurrence of pseudo-changes: In the fourth row of Figure 6, many methods fail to distinguish pseudo-changes on roads with appearances similar to buildings. In contrast, the proposed method can effectively identify the boundaries and main bodies of building changes. In the third row of Figure 7, areas near changed buildings exhibit different colors at different times, leading to pseudo-changes irrelevant to actual building changes. Many methods such as DMINet, STFF-GA, SEIFNet, and RaHFF cannot eliminate these pseudo-changes, while the method in this paper handles pseudo-changes caused by the diversity of land cover appearances effectively. In the first row of Figure 8, due to seasonal differences, roads show different colors at different times. Methods like BIT, ICIF-Net, DMINet, STFF-GA, and SEIFNet fail to identify these pseudo-changes, whereas the proposed method

demonstrates the best recognition effect for it. This advantage can be attributed to the TIEM method, which enhances temporal difference features based on a dual-branch path, reducing false detections.

- (2) Advantage in detecting edges and details: In the second row of Figure 6, the changing objects are small in area and densely arranged. Methods such as FC-Diff and DMINet have difficulty identifying details like the edges between densely clustered buildings. In the fourth row of Figure 7, methods such as DMINet and RaHFF struggle to recognize the differences between roads and buildings with similar colors. Although DMINet and RaHFF achieve relatively good change detection results, changes in the edge areas are easily overlooked.

In contrast, our DMTI-Net can effectively identify the boundaries and objects. These results can be attributed to the multi-scale context fusion capability of the LMCU and the inverse change map operation of the DM, which enhance the completeness of the content and improve the accuracy of boundary detection. In the visualization results of the LEVIR-CD dataset (Figure 6), to more prominently show that our method has the best detection performance for large-sized, small-sized, or densely clustered objects, we stitched 16 images of 256×256 into an image of 1024×1024 for display, demonstrating the overall optimal detection performance of our method.

- (3) Advantage in more complete detection of large objects: As shown in the first row of Figure 7, first and second rows of Figure 8, compared with the limitation of the comparison methods that are prone to miss detections in the recognition of complex semantic changes, this method can achieve a more complete change detection effect. This is mainly due to the following three aspects: Firstly, through the adjacent temporal feature fusion mechanism, the spatial semantic information in the temporal dimension is effectively integrated. Secondly, a multi-level temporal difference representation system is constructed, and the hierarchical feature capture mechanism is used to extract context features at different abstraction levels. Thirdly, based on the attention-guided feature refinement strategy, a progressive fusion method from global to local is adopted, and finally, a RSCD map with a fine edge structure is generated. This multi-level progressive fusion architecture ensures the complete perception of complex object changes in the RSCD task.

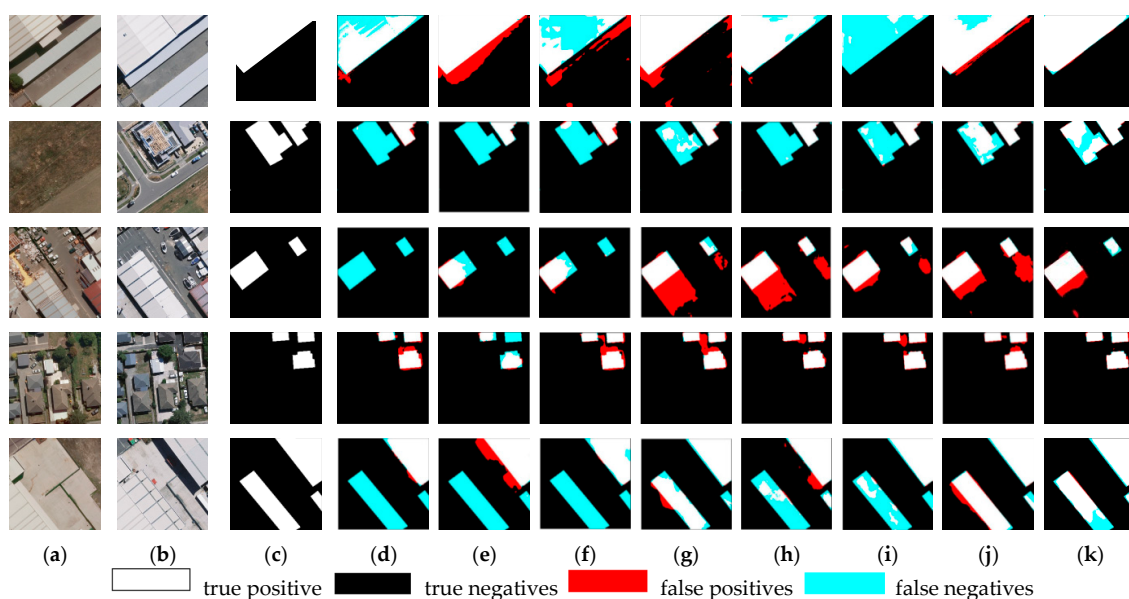


Figure 7. BCDD dataset. (a) t1 images (b) t2 images (c) Ground truth (d) FC-Diff (e) BIT (f) ICIF-Net (g) DMINet (h) STFF-GA (i) SEIFNetNet (j) RaHFF (k) Ours.

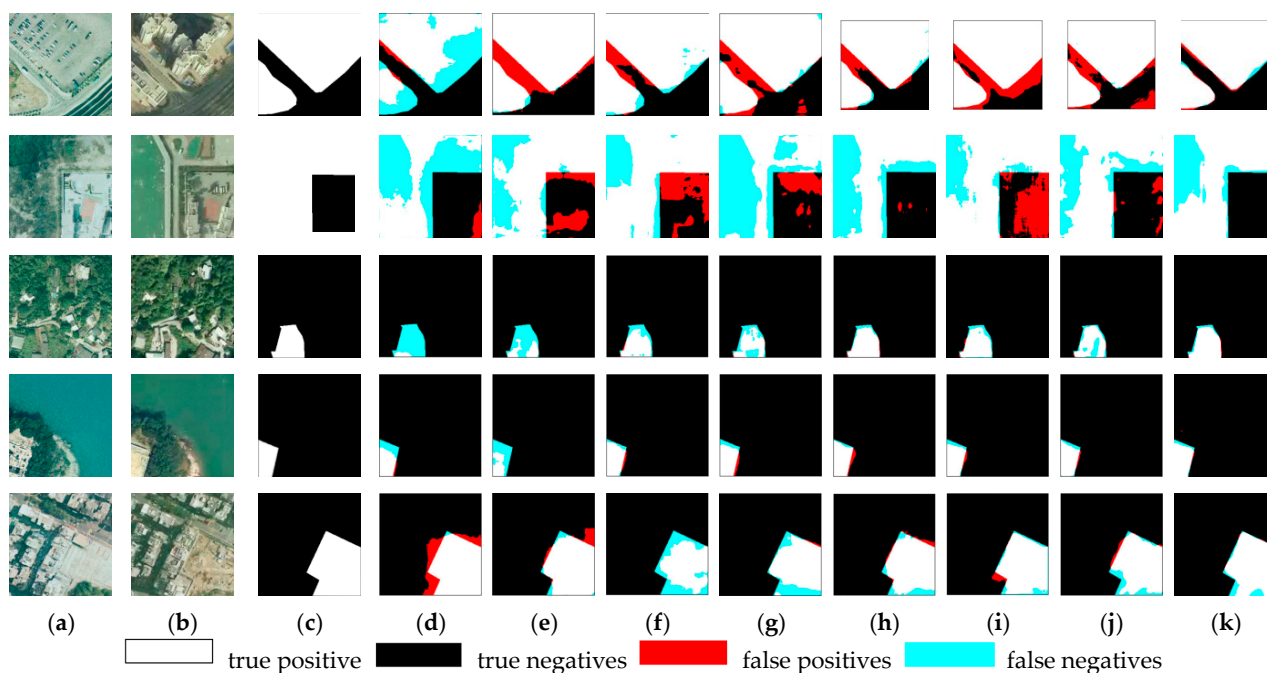


Figure 8. SYSU dataset. (a) t1 images (b) t2 images (c) Ground truth (d) FC-Diff (e) BIT (f) ICIF-Net (g) DMINet (h) STFF-GA (i) SEIFNetNet (j) RaHFF (k) Ours.

4.2.3. Comparison of Efficiency

In addition to the aforementioned quantitative and qualitative evaluation indicators, the efficiency of the model is also a crucial evaluation criterion. Efficiency is usually measured by the following two indicators: the number of model parameters (Params) and the number of floating-point operations (FLOPs). Figure 9 shows the comparison of the performance, parameters, and FLOPs of different models on the LEVIR-CD and SYSU-CD datasets. In the figure, the X-axis represents the efficiency indicators, and the Y-axis represents the evaluation accuracy indicator, where the F1 value is used here.

It can be seen from Figure 9 that models such as ICIF-Net, DMINet, and RaHFF, by introducing attention mechanisms or advanced feature extraction modules, have significantly improved the detection performance, although they have increased the overhead of the number of parameters and FLOPs. In contrast, models such as SEIFNet, STFF-GA, and BIT can still provide good detection accuracy while maintaining a low number of parameters and FLOPs.

Compared with these methods, our TIMA-Net model has a lower number of parameters and FLOPs than SEIFNet, STFF-GA, and BIT and at the same time performs excellently in terms of the F1 value, highlighting that while ensuring high detection accuracy, it has successfully balanced the detection performance and computational efficiency. This advantage is mainly attributed to the lightweight backbone network and the efficient module design. TIMA-Net employs MobileNetV2 as its backbone, which significantly reduces both the number of parameters and computational cost through depth-wise separable convolutions and linear bottleneck structures. Compared with networks based on conventional convolutions, MobileNetV2 offers higher computational efficiency and fewer parameters, thereby reducing the overall computational burden.

TIEM incorporates a lightweight spatiotemporal difference enhancement strategy, which effectively suppresses noise interference and focuses on changed regions through a dual-branch architecture and the Coordinate Attention (CA) mechanism. This module improves the accuracy of change detection while avoiding the redundant computation associated with subtraction-based methods in traditional approaches, further lowering

the computational overhead. Through the channel splitting design in LMCU, TIMA-Net can effectively aggregate multi-scale features without significantly increasing the computational cost.

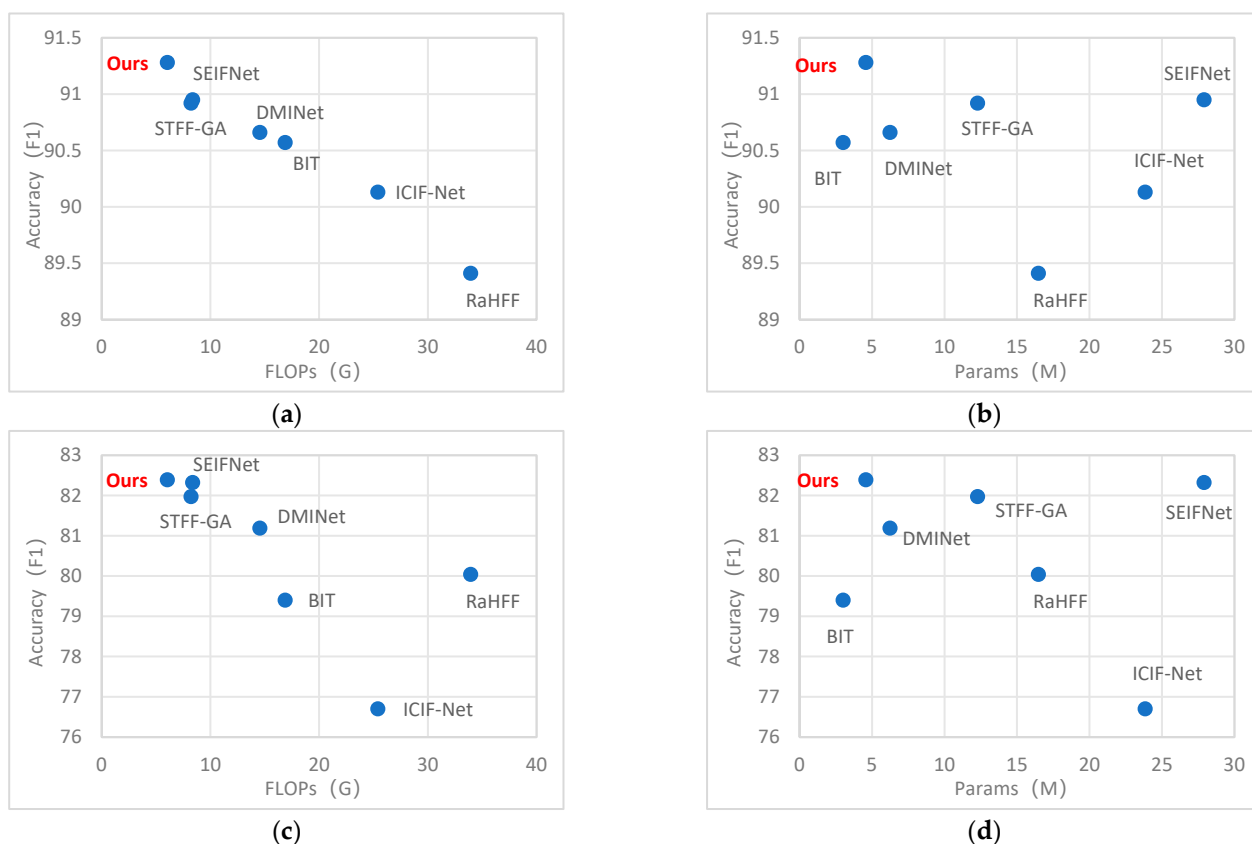


Figure 9. On the LEVIR and SYSU datasets, the model complexity of different methods in terms of Params (memory cost), FLOPs (computational cost), and F1 is compared. (a,b) are for the LEVIR dataset; (c,d) are for the SYSU dataset.

4.3. Ablation Experiments

To validate the effectiveness of the modules and components in the proposed AFRNet, we conducted a comprehensive ablation study on the LEVIR dataset.

(1) Effectiveness of CAFM

To overcome the limitations of the lightweight backbone in feature extraction, we designed the CAFM module, which enhances the representation capability of bi-temporal features by combining adjacent information extracted from different stages of the MobileNetV2 backbone. To verify the effectiveness of CAFM, we replaced it with a 3×3 convolution layer with unchanged input and output, labeled as “w/o CAFM” (Table 2, #02). The quantitative comparison results are shown in Table 2. The experimental results demonstrate that after removing CAFM, the IoU and F1 on the LEVIR-CD dataset decreased by 0.92% and 0.86%, respectively, and on the SYSU dataset, IoU and F1 decreased by 1.53% and 1.37%, respectively, indicating that the effective aggregation of multi-level features plays an important role in the TIMA-Net model.

Moreover, to systematically evaluate the contribution of the Feature Weighting (FW) strategy in the CAFM module for multi-level feature aggregation, we conducted an ablation experiment by removing the adaptive weighting mechanism (“CAFM w/o FW”, #03). The results demonstrate significant performance degradation across datasets; on LEVIR-CD, the IoU and F1-score decreased by 0.68% and 0.57%, respectively, while on SYSU, the declines

were more pronounced at 1.03% and 0.72%. This discrepancy is attributed to SYSU’s higher complexity (e.g., more diverse land cover types and finer change patterns), which amplifies the need for FW to prioritize discriminative features. Mechanistically, FW enables the model to dynamically suppress redundant low-level features (e.g., homogeneous background textures) and emphasize high-level change-related semantics (e.g., boundary details of construction sites).

Table 2. A quantitative comparison of IoU, F1, Rec, and Pre for different structural settings on two RSCD datasets was conducted.

NO.	Variants	LEVIR(%)				SYSU(%)			
		IoU	F1	Rec	Pre	IoU	F1	Rec	Pre
#01	TIMA-Net	83.96	91.28	89.97	92.62	70.06	82.39	80.49	84.38
Cross-Level Aggregation Fusion Module (CAFM)									
#02	w/o CAFM	83.04	90.42	88.95	91.93	68.53	81.02	79.11	83.02
#03	CAFM w/o FW	83.28	90.71	89.03	92.45	69.03	81.67	79.87	83.51
Temporal Interaction Enhancement Module (TIEM)									
#04	Sub	83.13	90.65	89.32	92.01	69.14	81.78	80.04	83.62
#05	Cat	82.95	90.41	88.93	91.94	68.71	81.28	79.42	83.25
#06	w/o DConv	83.31	90.89	89.84	91.98	69.18	81.91	80.35	83.52
Lightweight Multi-Scale Convolution Unit (LMCU)									
#07	w/o LMCU	82.98	90.42	88.96	91.93	68.14	80.66	78.67	82.76
Decoder Module (DM)									
#08	w/o DM	83.19	90.71	89.25	92.19	68.45	81.01	78.34	83.87
Loss Function									
#09	BCE	83.29	90.74	89.17	92.36	68.18	81.09	78.89	83.41
#10	Dice	83.54	90.93	89.46	92.45	68.85	81.71	79.24	84.33

We also directly compared CAFM with a similar strategy NAM (Table 3), and #02 in Table 4 represents the result of replacing our CAFM with NAM. The results show that compared with the NAM in the A2Net model, our CAFM can not only reduce the number of parameters but also has a better detection effect on the changing targets.

Table 3. A quantitative comparison using different feature aggregation modules on the LEVIR-CD dataset.

NO.	Modules	IoU	F1	Rec	Pre	FLOPs (G)	Params (M)
#01	CAFM(Ours)	83.96	91.28	89.97	92.62	6.05	4.31
#02	NAM(A2Net)	82.57	90.45	89.26	91.68	6.21	4.52

Table 4. A quantitative comparison of IoU, F1, Rec, and Pre for different Backbones on the LEVIR-CD dataset.

NO.	Backbones	IoU	F1	Rec	Pre	FLOPs (G)	Params (M)
#01	ResNet18	83.72	91.15	90.28	92.03	23.41	19.13
#02	EfficientNet	83.69	91.08	90.04	92.14	7.49	6.21

(2) Effectiveness of TIEM

TIMA-Net incorporates the TIEM module. To validate whether the TIEM module helps improve detection performance, we first replaced the TIEM module with a simple absolute subtraction operation, labeled as “Sub” (Table 2, #04), and reported the four evaluation metrics. The results show that after replacing TIEM with simple subtraction, the IoU and F1 on the LEVIR-CD dataset decreased by 0.83% and 0.63%, respectively, and on the SYSU-CD dataset, they decreased by 0.92% and 0.61%, respectively. Next, we replaced the TIEM module with channel-level concatenation, labeled as “Cat” (Table 2, #05), and obtained TIEMilar results showing performance degradation.

To further demonstrate the effectiveness of the TIEM design, we removed the dilated convolutions in the module and retained only the result of concatenation as the output of TIEM, labeled as “w/o DConv” (Table 2, #06). We found that after removing dilated convolutions, the IoU and F1 on the LEVIR-CD dataset decreased by 0.65% and 0.39%, respectively, and on the SYSU dataset, the IoU and F1 decreased by 0.78% and 0.48%, respectively. This shows that dilated convolutions can effectively search for fine-grained change features under different scale-receptive fields.

(3) Effectiveness of LMCU

To verify the effectiveness of the LMCU, we replaced it with a standard 3×3 convolution (denoted as w/o LMCU, #07 in Table 2). This substitution resulted in a decrease of 0.98% in IoU and 0.86% in F1 score on the LEVIR-CD dataset and a decrease of 1.92% in IoU and 1.71% in F1 score on the SYSU dataset. These results indicate that the multi-scale convolutions in different directions within the LMCU contribute significantly to the overall performance improvement of the module.

(4) Effectiveness of the DM

The DM module uses a self-attention mechanism to weight the feature maps and enhance the recognition of change regions. To validate the effectiveness of the DM module, we replaced it with a regular convolution layer, denoted as “w/o DM” (Table 2, #08). Quantitative results showed that after removing the DM module, the IoU on the LEVIR-CD dataset decreased by 0.77% and F1 decreased by 0.57%, indicating that the weighting of feature maps by the DM module plays an important role in improving change detection accuracy. This result suggests that the self-attention mechanism effectively helps the model focus on the change regions, thereby further improving the accuracy of change detection.

(5) Effectiveness of the Loss Function

To study the impact of the loss function on performance, we used traditional binary cross-entropy (BCE) loss and Dice loss as baselines and compared them with the hybrid loss (BCE + Dice). The results are reported in Table 2, #09 (BCE) and #11 (Dice). The TIMA-Net model using the hybrid loss (Table 2, #10) outperformed those using only BCE loss or Dice loss. These results indicate that the hybrid loss function significantly improves the model’s performance, especially in handling imbalanced classes, as it better balances the contributions of each class and enhances detection accuracy.

(6) Comparison of Different Backbone Networks

To verify the rationality of selecting the backbone network, we compared the performances of ResNet18, EfficientNet, and MobileNetV2 on the LEVIR-CD dataset (Table 4). The results show that although the model using MobileNetV2 as the backbone network is not optimal in terms of the Recall metric, its F1 and IoU metrics outperform those of ResNet18 and EfficientNet. Moreover, in terms of FLOPs and Params, it has a significant advantage over ResNet18 (23.41 G, 19.13 M) and EfficientNet (7.49 G, 6.21 M) in terms

of computational cost and parameter count. Considering the lightweight positioning of TIMA-Net, MobileNetV2, with its depth-wise separable convolution and linear bottleneck structure, balances feature extraction and computational efficiency, making it more suitable for the actual deployment requirements of remote sensing change detection.

5. Discussion

We propose a lightweight remote sensing image change detection network, TIMA-Net, aiming to improve the accuracy and computational efficiency of change detection in high-resolution remote sensing images. By introducing the Cross-Level Feature Aggregation Module (CAFm), the Spatio-Temporal Interaction Enhancement Module (TIEM), and the Lightweight Multi-Scale Convolution Unit (LMCU), TIMA-Net can effectively capture subtle changes in remote sensing images while reducing computational costs and adapting to resource-constrained application environments.

Specifically, TIMA-Net enhances the representation ability of change regions through multi-scale feature fusion and spatio-temporal difference enhancement techniques and shows strong robustness, especially in capturing complex change regions and details. Ablation experiments further verify the contribution of each module to the final performance, demonstrating the crucial role of adaptive dilated convolution and spatio-temporal difference enhancement strategies in reducing noise and improving detection accuracy.

Although this study has achieved satisfactory results in the task of change detection of high-resolution remote sensing images, one of the key compromises related to lightweight networks is the potential trade-off between computational efficiency and feature extraction ability. Although TIMA-Net is designed to reduce computing costs and memory usage, this may lead to the loss of some fine-grained features, especially in the field of highly complex changes.

6. Conclusions

This study presents TIMA-Net, a lightweight remote sensing image change detection network designed to balance accuracy and computational efficiency in high-resolution imagery. By integrating the Cross-Level Feature Aggregation Module (CAFm), Spatio-Temporal Interaction Enhancement Module (TIEM), and Lightweight Multi-Scale Convolution Unit (LMCU), TIMA-Net achieves efficient feature fusion and noise suppression. Experimental results on LEVIR, SYSU, and BCDD datasets demonstrate its superiority, with an F1-score of 91.28% on LEVIR-CD using only 4.31 M parameters—significantly outperforming state-of-the-art methods in both accuracy and computational efficiency. Ablation studies validate that each module contributes to improved performance, particularly CAFm's reduction in complexity and TIEM's joint metric of Euclidean distance and cosine similarity, which enhances resistance to seasonal pseudo-changes.

TIMA-Net's implications lie in its practical utility for resource-constrained scenarios, such as UAV-based monitoring or edge device deployment, where traditional heavyweight networks are infeasible. However, the lightweight design introduces trade-offs; while it efficiently captures macro-scale changes, fine-grained details in highly complex scenes (e.g., small infrastructure modifications) may be underrepresented, when processing densely clustered objects. This highlights the need for balanced feature extraction that preserves both semantic context and spatial detail.

Future research will focus on further reducing the number of parameters of the model while improving the accuracy. Specifically, by optimizing the network structure and adopting more efficient feature extraction methods and quantization techniques, the computational overhead of the model can be further reduced. In addition, combining the adaptive learning mechanism with new optimization algorithms will help improve the ac-

curacy and robustness of the model in complex scenarios and further enhance the practical application potential of the model.

Author Contributions: Conceptualization, Z.Z.; Methodology, Z.Z.; Validation, X.L.; Formal analysis, L.W. (Lvchun Wang); Data curation, X.Z. and W.Y.; Writing—original draft, Z.Z.; Writing—review & editing, Z.Z., X.Z. and L.W. (Longbao Wang); Visualization, Z.Z. and S.X.; Funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The LEVIR-CD dataset can be found at <https://justchenhao.github.io/LEVIR> (accessed on 1 December 2022). The SYSU dataset can be found at <https://github.com/liumency/SYSU-CD> (accessed on 1 December 2022). The BCDD dataset can be found at http://study.rsgis.whu.edu.cn/pages/download/building_dataset.html (accessed on 1 December 2022). Our source code will be published on <https://github.com/xiangxianghaochi/TIMA-Net> (accessed on 1 December 2022).

Conflicts of Interest: Authors Xiaoliang Luo, Lvchun Wang and Wei Yu were employed by the company Jiangxi Virtual Reality Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- de Bem, P.P.; de Carvalho Júnior, O.A.; Guimarães, R.F.; Gomes, R.A.T. Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks. *Remote Sens.* **2020**, *12*, 901. [CrossRef]
- Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sens. Environ.* **2021**, *265*, 112636–112657. [CrossRef]
- Chen, J.; Liu, H.; Hou, J.; Yang, M.; Deng, M. Improving building change detection in VHR remote sensing imagery by combining coarse location and co-segmentation. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 213. [CrossRef]
- Ke, L.; Lin, Y.; Zeng, Z.; Zhang, L.; Meng, L. Adaptive change detection with significance test. *IEEE Access* **2018**, *6*, 27442–27450. [CrossRef]
- Coppin, P.R.; Bauer, M.E. Digital change detection in forest ecosystems with remote sensing imagery. *Remote Sens. Rev.* **1996**, *13*, 207–234. [CrossRef]
- Lambin, E.F.; Strahlers, A.H. Change-vector analysis in multitemporal space: A tool to detect and categorize land-cover change processes using high temporal-resolution satellite data. *Remote Sens. Environ.* **1994**, *48*, 231–244. [CrossRef]
- Celik, T. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [CrossRef]
- Crist, E.P. A TM tasseled cap equivalent transformation for reflectance factor data. *Remote Sens. Environ.* **1985**, *17*, 301–306. [CrossRef]
- Gong, J.; Hu, X.; Pang, S.; Li, K. Patch matching and dense CRF-based co-refinement for building change detection from bi-temporal aerial images. *Sensors* **2019**, *19*, 1557. [CrossRef]
- Im, J.; Jensen, J. A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sens. Environ.* **2005**, *99*, 326–340. [CrossRef]
- Verdier, G.; Ferreira, A. Adaptive Mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing. *IEEE Trans. Semicond. Manuf.* **2011**, *24*, 59–68. [CrossRef]
- Seo, D.K.; Kim, Y.H.; Eo, Y.D.; Lee, M.H.; Park, W.Y. Fusion of SAR and multispectral images using random forest regression for change detection. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 401. [CrossRef]
- Jiang, W.; Sun, Y.; Lei, L.; Kuang, G.; Ji, K. Change detection of multisource remote sensing images: A review. *Int. J. Digit. Earth* **2024**, *17*, 2398051. [CrossRef]
- Li, X.; He, M.; Li, H.; Shen, H. A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
- Li, X.; Du, Z.; Huang, Y.; Tan, Z. A deep translation (GAN) based change detection network for optical and SAR remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *179*, 14–34. [CrossRef]
- Bai, B.; Fu, W.; Lu, T.; Li, S. Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]

17. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [[CrossRef](#)]
18. Peng, D.; Zhang, Y.; Guan, H. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [[CrossRef](#)]
19. Zhu, Q. Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 63–78. [[CrossRef](#)]
20. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
21. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
22. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
23. Fang, S.; Li, K.; Li, Z. Changer: Feature interaction is what you need for change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11. [[CrossRef](#)]
24. Liang, S.; Hua, Z.; Li, J. Enhanced feature interaction network for remote sensing change detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–15. [[CrossRef](#)]
25. Sun, Y.; Lei, L.; Li, Z.; Kuang, G. Similarity and dissimilarity relationships based graphs for multimodal change detection. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 70–88. [[CrossRef](#)]
26. Zhang, H.; Lin, M.; Yang, G.; Zhang, L. ESCNet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 28–42. [[CrossRef](#)]
27. Li, Q.; Zhong, R.; Du, X.; Du, Y. TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [[CrossRef](#)]
28. Daudt, R.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing, ICIP, Athens, Greece, 7–10 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4063–4067.
29. Zhang, Z.; Wang, X.; Jung, C. DCSR: Dilated convolutions for single image super-resolution. *IEEE Trans. Image Process.* **2018**, *28*, 1625–1635. [[CrossRef](#)]
30. Feng, Y.; Xu, H.; Jiang, J.; Liu, H.; Zheng, J. ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [[CrossRef](#)]
31. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure transformer network for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [[CrossRef](#)]
32. Yan, L.; Jiang, J. A hybrid siamese network with spatiotemporal enhancement and two-level feature fusion for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [[CrossRef](#)]
33. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 18–22 June 2018; pp. 6848–6856.
34. Li, Z.; Tang, C.; Liu, X.; Zhang, W.; Dou, J.; Wang, L.; Zomaya, A.Y. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. [[CrossRef](#)]
35. Lei, T.; Geng, X.; Ning, H.; Lv, Z.; Gong, M.; Jin, Y.; Nandi, A.K. Ultralightweight spatial-spectral feature cooperation network for change detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [[CrossRef](#)]
36. Codegoni, A.; Lombardi, G.; Ferrari, A. TINYCD: A (not so) deep learning model for change detection. *Neural Comput. Appl.* **2023**, *35*, 8471–8486. [[CrossRef](#)]
37. Huang, Y.; Li, X.; Du, Z.; Shen, H. Spatiotemporal Enhancement and Interlevel Fusion Network for Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [[CrossRef](#)]
38. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13708–13717.
39. Milletari, F.; Navab, N.; Ahmadi, S. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
40. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [[CrossRef](#)]
41. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [[CrossRef](#)]

42. Feng, Y.; Jiang, J.; Xu, H.; Zheng, J. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [[CrossRef](#)]
43. Wei, H. Spatio-Temporal Feature Fusion and Guide Aggregation Network for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642216. [[CrossRef](#)]
44. Wang, B.; Zhao, K.; Xiao, T.; Qin, P.; Zeng, J. RaHFF-Net: Recall-Adjustable Hierarchical Feature Fusion Network for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 176–190. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.