TIMF-Net: Temporal Interaction and Multiscale Fusion Networks for Remote Sensing Change Detection

Shuo Wang, Zhiqing Zheng, and Jinjiang Li

***Abstract*—In recent years, the field of remote sensing change detection (RSCD) has experienced transformative advancements through the application of convolutional neural networks (CNNs). However, inconsistencies in image quality, noise, and pseudochanges caused by variations in illumination, climate, and surface conditions due to different acquisition times pose significant challenges. Addressing these issues, this study increases traditional RSCD methodologies by introducing a novel temporal interaction and multiscale fusion network (TIMF-Net). TIMF-Net incorporates a temporal interaction and difference enhancement module (TIDEM) that effectively extracts and augments change information within images. This module deeply integrates temporal information through a weighted fusion strategy, not only capturing the juxtaposition and superposition relationships between images but also unraveling complex feature representations to ensure accurate alignment and coupling of features across different periods. Additionally, we propose a multiscale global-aware (MSGA) module, which attends to both local details and global contextual information, integrating pixel-level features and demonstrating heightened sensitivity to multiscale changes such as path alterations, water fluctuations, and agricultural variations. TIMF-Net outperforms mainstream and state-of-the-art methods on three datasets, achieving an F1 score of 91.96% and intersection over union (IoU) of 85.12% on the LEVIR-CD dataset, an F1 of 93.37% and IoU of 87.56% on the WHU-CD dataset, and an F1 of 87.12% and IoU of 77.19% on the GZ-CD dataset, with 27.64 M Params and 42.8 G FLOPs.**

***Index Terms*—Attention mechanism Transformer, building change detection (BCD), dual graph-convolution module, high-resolution remote sensing (RS) images.**

I. Introduction

BITEMPORAL remote sensing change detection (RSCD) is a pivotal concept within the discipline of remote sensing, involving the analysis of satellite imagery captured over a specific terrestrial area at two distinct temporal points. The crux of this concept rests on the utilization of paired remote sensing imagery—constituting a time series dataset comprising a baseline (or reference) image and a subsequent comparison image. The baseline image embodies the state of the region at the initial time point, whereas the comparison image reflects the status at a later time point. Through the comparative analysis of these two images, discernible variations in land use [1], [2], [3], [4], vegetation cover [5], urban expansion [6], [7], aquatic changes [8], impacts of disasters [9], [10], and shifts in ecological environments can be identified and quantified.

RSCD constitutes a pivotal analysis of satellite imagery captured over a specific terrestrial domain at two distinct time points. This analytical approach seeks to output a binary image delineating the changed regions [11], as depicted in Fig. 1. It transcends mere segmentation, engaging in a sophisticated process of feature identification. Change detection (CD) inherently deals with multiple input images, escalating computational demand exponentially with image quantity, hence the computational load typically exceeds that of simple image segmentation tasks. Moreover, discerning pivotal from negligible features is crucial when analyzing temporal remote sensing images, particularly as factors like vegetation growth, seasonal transitions, or lighting variations could induce disparities in feature appearance. Accurate detection hinges on identifying and concentrating on truly pertinent features.

RSCD encounters numerous challenges. Precise spatial registration is required for accurate alignment of images from different periods, compounded by the need for meticulous radiometric correction to account for radiometric discrepancies, thus ensuring the reliability of analysis. Furthermore, the complexity and variability of land surface features, due to diverse natural environments and anthropogenic activities, add to the difficulty of identifying genuine changes. The phenomenon of pseudochanges warrants vigilance; as seasonal shifts, atmospheric conditions, cloud cover, or inherent image noise could all mislead change signals. Thus, the selection and extraction of suitable features are critical, directly influencing the accuracy and efficiency of CD. These challenges necessitate continual solutions post-image correction, with new CD algorithms being proposed and applied over decades. Initially constrained by computational hardware capabilities, such as GPUs, CD tasks

Manuscript received 23 May 2024; revised 6 July 2024; accepted 25 July 2024. Date of publication 30 July 2024; date of current version 9 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61772319, Grant 62002200, Grant 62202268, and Grant 62272281, in part by Shandong Natural Science Foundation of China under Grant ZR2023MF026 and Grant ZR2022MA076, and in part by Yantai Science and Technology Innovation Development Plan under Grant 2022JCYJ031 and Grant 2023JCYJ040. *(Corresponding author: Zhiqing Zheng.)*

Shuo Wang and Jinjiang Li are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China.

Zhiqing Zheng is with the Department of Library Information, Shandong Technology and Business University, Yantai 264005, China (e-mail: 201511652@sdtbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3435575

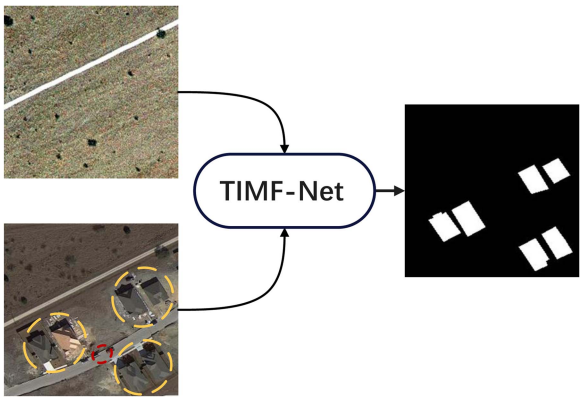

Fig. 1. Flowchart of RSCD in dual time zones CD. Satellite images taken at two different time points are fed into the TIMF-Net network to generate a binary change map. The unchanged areas are shown in black and the changed areas are shown in white. Yellow circles indicate detected changes and red circles indicate unchanged features in the scene. The main purpose here is to detect changes in buildings.

predominantly utilized traditional machine learning methods like random forests [12], decision trees [13], and support vector machines [14] for bitemporal image CD and pixel classification, generating CD results. While these traditional methods have realized significant achievements across multiple domains, they exhibit evident weaknesses in handling complex targets and environmental noise within images, especially limited in accuracy for target recognition and detection. Nonetheless, these early studies laid the groundwork for subsequent breakthroughs with deep learning technology.

The advent of deep learning technology has significantly advanced remote sensing image processing. Previously, CD relied heavily on algebraic operations to analyze data differences, suitable for low-resolution data. With technological progression, CNN-based methods began to supplant traditional algorithms [15], [16], exhibiting immense potential, particularly in processing high-resolution images and excelling in capturing rich and abstract local contextual features, thus becoming a driving force for domain exploration [17]. Beyond CNNs, other models like GANs [18] and GCNs [19] have entered the realm of CD, optimizing data modeling and feature extraction. Some excellent ways of mixing pixels are also shown [20].There are also many advanced deep learning methods [21], [22], [23].These algorithms, with their stellar performance, have promoted end-to-end change recognition in remote sensing data, streamlining the processing workflow. Deep learning has not only achieved tremendous strides in algorithmic innovation but also displayed characteristics of low resource occupancy in resource utilization [24], [25], [26].

With the evolution of Transformer models [27], their unique self-attention mechanism plays a pivotal role in RSCD. Capable of capturing global contextual information, Transformers comprehend interpixel relationships across the entire image [28], [29], enabling significant performance enhancements in processing multiscale features and fusing multitemporal remote sensing data. Transformers have also refined the model's capability to discern boundaries of change areas. For instance, BIT [30] has enhanced CD accuracy and boundary precision via Transformers, augmenting semantic feature information. Their flexible architecture allows for integration with other technologies, suited for monitoring complex and dynamically changing terrestrial environments, set to transform traditional RSCD methods, offering more efficient and accurate solutions.

Current CD approaches often emphasize feature extraction from single images or rudimentary differential feature extraction [31], neglecting the spatial feature interaction of bitemporal change images, leading to some information wastage and feature loss. Moreover, when contending with large-scale scene changes and disturbances such as lighting discrepancies, the reliability, robustness, and multilevel difference capture capacity for detecting various scale targets require comprehensive consideration. Hence, an efficient and effective model is necessitated, one that consistently focuses on the capture of differential and scale-diverse feature characteristics. Feature interaction facilitates the propagation of information between bitemporal images, aiding the shared spatial information. At last, considering the imbalance conundrum between foreground and background categories, especially on local and global feature information, it becomes imperative to concentrate attention on mutually guiding change areas. By propagating local and global features of bitemporal images to one another, the model can regulate attention distribution while maintaining its features, thereby better managing spatiotemporal feature differences.

In order to solve the above problems, we propose a new temporal interaction and multiscale fusion network (TIMF-Net) network, as shown in Fig. 2. Specifically, we used PVTv2 [32] for single image feature extraction, and the PVTv2 model improves the robustness and effectiveness of the model by virtue of its multiscale feature fusion, optimized computational efficiency, and the attention mechanism, which is much better compared to the original Transformer's superior performance and wider application potential, and also surpasses the traditional CNN [33] model in global context understanding and feature capture. In addition, we propose a novel feature interaction fusion method: temporal interaction and difference enhancement module (TIDEM). TIDEM not only considers the superposition and difference information between two feature maps but also incorporates the multiplicative relationship and optimal value information between them. Inspired by the need for temporal data processing, particularly the surface changes observed in remote sensing imagery, TIDEM enhances the recognition accuracy and efficiency of remote sensing imagery under complex environmental conditions by reinforcing multiple feature interactions between time points. For example, the multiplication operation helps to reinforce persistent significant changes, while the maximum operation helps to resist episodic noise or environmental disturbances. The combined use of these methods improves detection accuracy and reduces false alarms and missed detections. Furthermore, TIDEM enhances the difference of the four feature fusions after global and local feature extraction and

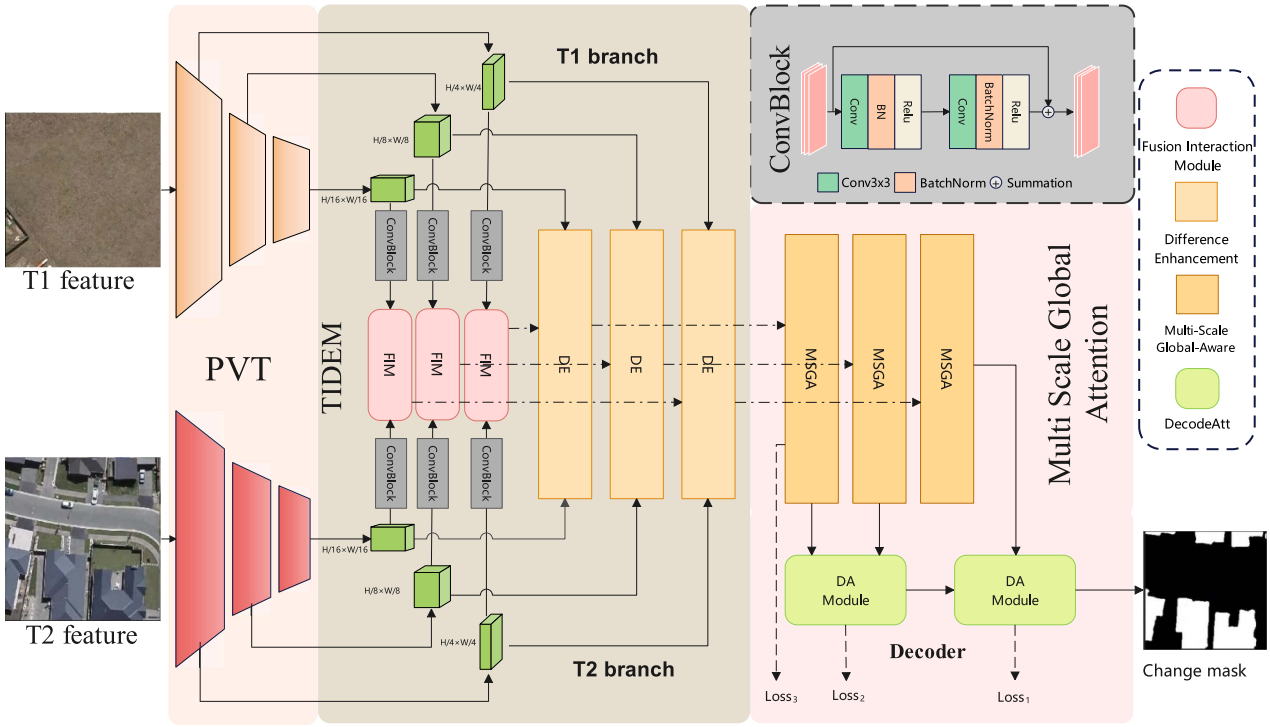

Fig. 2. Illustrates the proposed method framework. Utilizing the PVTv2-b1 algorithm, basic features are extracted from a pair of remote sensing images. Subsequently, at the corresponding feature levels, TIDEM module engages in feature temporal interaction and enhancement. The enhanced features are then captured by the MSGA module to incorporate multiscale feature information, thereby improving the accuracy of CD. Finally, the refined change map is obtained through polishing by the DA module.

redistributes the weights of dual-time image features, allowing the model to utilize complete information from different time points rather than relying on a single data source. This approach increases the information available for decision-making and helps form a more comprehensive judgment of change. Subsequently, we designed a multiscale global perception module, multiscale global-aware (MSGA), addressing the need for multiscale and detailed feature representation in remote sensing images. MSGA combines pixel-level attention and multiscale channel attention, emphasizing the local and global parts of the image regarding change and subtle change capture. The channel attention mechanism further refines the feature weights at each scale, enhancing the network's ability to understand and process important features at the pixel level. Finally, the proposed decoding attention (DA) module strengthens the model's ability to perceive critical change regions. The introduction of the attention mechanism at the decoding stage enhances the model's sensitivity to spatial features, making it more accurate and effective in recovering image details, especially when dealing with fine and complex targets. In summary, the main contributions are as follows.

1) We have developed a module named TIDEM, which enhances feature interaction through various operations, significantly improving the detection capabilities and accuracy for terrestrial changes. Ultimately, by stacking difference enhancement and attention-guided strategies, we have achieved precise and effective recognition of change information in remote sensing imagery.
2) We introduced an MSGA module, which, through the extraction of interactive information across channels and scales, effectively captures key features at different detail levels within the image. Furthermore, through a pixel-level attention mechanism, this module models and focuses on the significant pixels involved in changes within remote sensing imagery.
3) We designed a DA module that notably enhances the spatial feature discrimination and target detail restoration capabilities during the crucial decoding phase.

The rest of this article is organized as follows. Section II briefly reviews related work. The details of our proposed framework are described in Section III. Comprehensive experimental evaluations are conducted in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Traditional CD Methods

Over the past few decades, CD has rapidly advanced to meet the diverse demands of various application scenarios. This evolution has seen a transition from basic algorithmic approaches to more sophisticated machine learning methods, representing significant progress in the field. The primary task of RSCD involves analyzing disparities between remote sensing images captured at different points in time to discern terrestrial alterations. Typically, this process entails comparing two images, often through direct subtraction, and then determining areas of

Algorithm 1: TIMF-Net.

Input: X_1,X_2.*(bitemporal image)*
Output: O.
// Step1 : Use the PVTv2-b1 to fuse multiscale features in the encoder
for i in 1,2 **do**
  for i in 1,2,3 **do**
    $X_{i,j} = \text{PVT}(X_i)$;
  end
end
// Step2 : Use the TIDEM module to Calculate the feature Interaction
for i in 1,2,3 **do**
  $X_m = \text{TIDEM}(X_{m,1}, X_{m,2})$;
  // Step3 : Use the MSGA module
  $X'_m = \text{MSGA}(X_{m,1}, X_{m,2})$;
end
// Step4 : Use the DA module
$X_n = \text{DA}(X'_2, X'_3)$;
$X'_n = \text{DA}(X'_1, X_n)$;
// Step5 : obtain the result
$\mathbf{O} = \text{Conv}(X'_n)$

change based on the variance in pixel values within the resultant differential image, typically set by predefined thresholds. For instance, Mahmoudzadeh [34] utilized threshold-based CD methods, wherein they compared remote sensing images from two different time periods and evaluated pixel value changes based on predetermined thresholds. Coppin et al. [35] employed spectral and spatial information extracted from remote sensing images, such as texture and morphology, in conjunction with machine learning algorithms or empirical models to identify changes. Similarly, Swain and Davis [36] conducted land cover change analysis using multispectral remote sensing data, employing image classification to compare images from different time periods and subsequently discern changes based on variations in classification results. While these traditional methods have been essential in RSCD research, they often fell short in fully leveraging the complex information inherent in remote sensing data, leading to inadequate accuracy and robustness in CD.

With time, the application of machine learning technology in RSCD has become increasingly widespread, significantly enhancing the automation and accuracy of CD. For instance, Bovolo and Bruzzone [37] applied support vector machines for unsupervised CD in large-area multitemporal images, effectively assessing the damage caused by tsunamis. Gómez et al. [38] utilized the random forest method for land cover classification of optical remote sensing time series data. Liu and Chen [39] classified feature vectors using the K-nearest neighbors algorithm to determine the change status of each pixel, evaluating the algorithm's performance by comparison with actual change images. Despite the efficiency and accuracy improvement of machine learning technology [40] in the remote sensing field, challenges such as insufficient generalization ability, poor model interpretability, high sensitivity to data quality and feature selection, and dependence on substantial computational resources remain.

B. CNN Methods

Since their introduction, convolutional neural networks (CNNs) [41] have demonstrated their powerful capability in image processing and computer vision fields, particularly in feature extraction. CNNs, with their unique convolutional layer structure, can effectively learn useful feature representations automatically from raw images, which is particularly important in processing complex remote sensing imagery. The complexity of remote sensing images, including but not limited to their high dimensionality and rich geographical information features, necessitates a powerful tool capable of capturing and understanding this complexity, for which CNNs have been widely applied [42].

In RSCD, CNN feature fusion technology has proven its significance in enhancing detection accuracy and efficiency. CNN feature fusion integrates data from different time points, sensors, or scales of remote sensing images, deeply mining and utilizing the complex spatial and spectral information contained within these data. For example, processing multitemporal images with CNNs can effectively capture subtle features of terrestrial changes over time, thereby improving CD accuracy [43]. Furthermore, fusing features from optical and SAR images not only enhances the model's ability to recognize different types of land cover but also improves detection performance under complex environmental conditions (e.g., cloud coverage) [44]. Multiscale feature fusion techniques, such as integrating deep and shallow features in CNNs, further enhance the capture of surface detail changes [45], allowing the model to understand both the global structure and local details of images. Recent advancements in deep learning have introduced new feature fusion strategies, such as attention-based methods, focusing the network more on regions with significant changes [46], [47], thus, improving the accuracy and robustness of CD tasks. These advancements not only propel the development of RSCD technology but also offer new directions and methods for future research in remote sensing image processing and analysis. Despite significant progress in current research in handling multitemporal remote sensing data, where images from different time points contain crucial clues for change information, existing methods have not effectively fused information from these different time points, leading to inaccurate recognition or missed detection of change areas.

C. Transformer-Based Models

The application of the Transformer model in the field of RSCD has made significant progress in recent years. Its self-attentive mechanism is widely recognized as particularly effective in capturing long-range dependencies in image sequences, which is especially important for analyzing remotely sensed data over time [48]. The Transformer model shows great potential in CD tasks due to its high efficiency in processing remotely sensed data and its sensitivity to subtle temporal changes. Yuan [49] et al., by combining a novel network with UNet architecture

and Transformer model, this combined approach exploits the powerful feature extraction capability of UNet and the global self-attention mechanism of Transformer in order to overcome the limitations of traditional CNN networks in modeling global dependencies. Xu et al. [50], based on the Transformer's remote sensing image CD method, significantly improved the ability to distinguish pseudochanges and the overall detection accuracy by introducing visual tokenization, progressive visual Transformer, and multimodule fusion. In addition, efficient Transformer variants are being developed to adapt to the processing needs of large-scale remote sensing data and reduce the computational cost while maintaining high detection performance.

In recent years, the fusion of CNNs and Transformers has demonstrated advanced performance in RSCD tasks. For example, Chen et al. [30] introduced the BIT network, which incorporates Transformer encoders and decoders for modeling spatiotemporal context and feedback into the pixel space. Specifically, the BIT module first utilizes a Transformer encoder to model spatiotemporal context based on abstract feature mappings from CNNs. In another study, Feng et al.'s [51] ICIF-Net combined the strengths of CNNs and Transformers through internal scale-cross interaction and interscale feature fusion, effectively capturing local and global features and addressing alignment issues caused by downsampling operations in traditional networks.

III. Methodology

A. Framework Overview

At the core of our study, we meticulously elaborate on the network's key architectures, including the TIDEM module, the MSGA module, and the DA module, as illustrated in Fig. 2. The details are as follows.

We have selected the PVTv2-B1 from the Transformer architecture as our core backbone network, owing to the PVTv2's design philosophy highly aligning with the requirements of CD tasks: the sensitive capture of subtle but crucial changes, effective handling of objects of various scales, and robustness to environmental changes (e.g., lighting variations, seasonal transitions). Specifically, for PVTv2-B1, we applied multiscale feature mappings with channel numbers of 64×64, 128×128, and 320×320 in its first three stages, providing a solid foundation for capturing and processing change phenomena across various scales. Building upon this foundation, we employ the temporal interaction and difference enhancement (TIDEM) module to precisely process and fuse the differences in bitemporal features at different levels. The coupled features processed by TIDEM are then fed into the MSGA module, where features at multiple scales are comprehensively enhanced and contextually adjusted, while retaining key geographic and environmental information. Finally, through the DA module, we process and integrate information from different levels, facilitating interactive processing of the encoder input feature maps. The learning process is then supervised by a deep supervision mechanism, specifically, each deep supervision layer outputs a result processed through upsampling and convolution, culminating in an accurate prediction image generated via argmax operation.

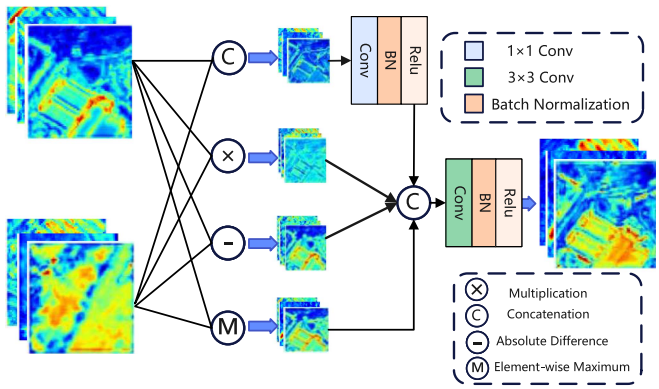

Fig. 3. Illustrates the fusion interaction process. T1 and T2 represent feature images from different time points, and Fusion denotes the result of the fusion interaction.

B. Temporal Interaction and Difference Enhancement Module

In RSCD tasks, relying solely on simple difference operations or dimensional concatenation often results in interference from seasonal changes or lighting effects, making it difficult to accurately capture subtle changes. Existing time interaction methods also do not fully utilize the temporal dimension information. Considering the influence of irrelevant factors such as atmospheric conditions and cloud cover when calculating difference images using bitemporal images, which may disrupt the numerical changes in the image pixels and affect the analysis of differences between the two images, we propose a time interaction and difference enhancement framework incorporating a fusion interaction module (FIM) and a difference enhancement (DE) module. This framework enhances the diversity and expressiveness of features through four interaction techniques. The first method involves channel stacking to capture features common to both the baseline and subsequent images, allowing models or algorithms to access the complete data from both time points for a more comprehensive analysis of change. The second method uses multiplication operations, It naturally suppress static or unchanged background information by producing relatively small products for areas without significant changes. The third method employs absolute value subtraction to highlight areas with substantial changes, ensuring that regions with only minor changes are naturally suppressed. The fourth method utilizes a maximum value operation, It can resist disturbances caused by shadows or uneven lighting. These factors typically do not produce high values in images, so the maximum value selection naturally ignores these lower readings. Additionally, important features such as the edges of buildings or the boundaries of water bodies are often more prominent due to higher reflectance. By fusing these four types of features, we enhance the complementarity of the features. As shown in Fig. 3, we use a heatmap to vividly demonstrate the results of each interaction operation in the shallow features, providing a more intuitive display of how the DE module can perform feature extraction and enhancement, Following this, global and local feature weights are derived and assigned to both the baseline and subsequent images to suppress

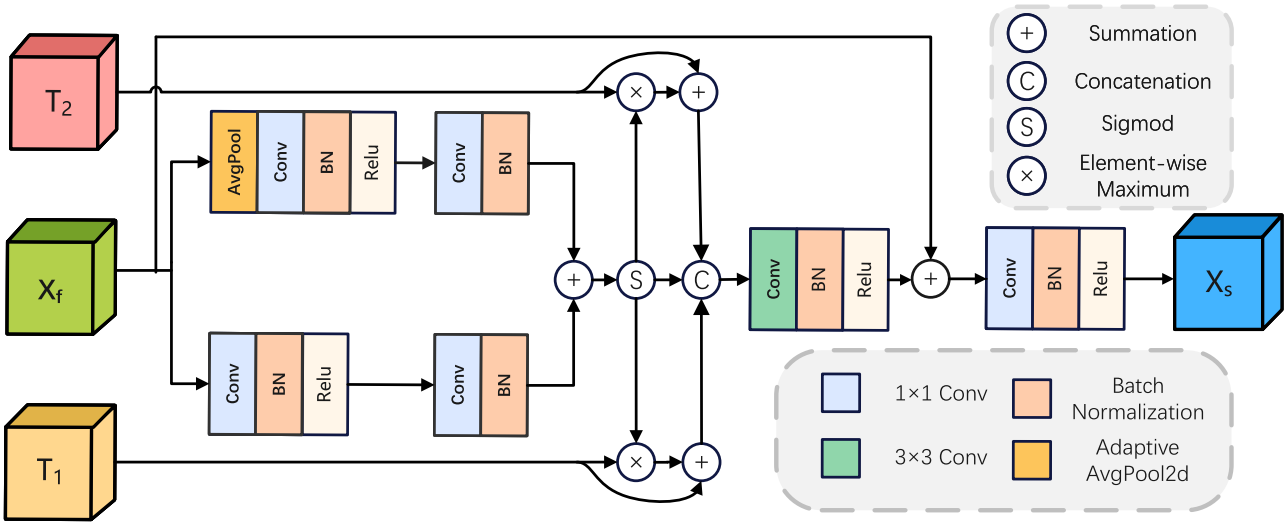

Fig. 4. Illustrates the process of the difference enhancement module.

irrelevant information and highlight change features, thereby enhancing performance in complex environments, as illustrated in Fig. 4.

This module takes the scene features captured at two distinct time points $T_1 \in \mathbb{R}^{H\times W\times C}$ and $T_2 \in \mathbb{R}^{H\times W\times C}$as input, where C denotes the number of channels and H and W denote the height and width, respectively. First, each scale feature T_1 and T_2 of the PVT output is processed twice through a 3×3 convolution module for shallow feature deepening, as shown in the ConvBlock of Fig. 2 After processing, the outputs of the convolutional layers are connected back to the original inputs T_1 and T_2 via residual concatenation (where R denotes ReLU and B denotes BatchNorm2d), which can be articulated by the following, resulting in the enhanced features $X_1 \in \mathbb{R}^{H\times W\times C}$ and $X_2 \in \mathbb{R}^{H\times W\times C}$:

$$\begin{aligned} X_1 &= R\left(B\left(\text{Conv}\,3\times 3\left(T_1\right)\right)\right)_{\times 2} \\ X_2 &= R\left(B\left(\text{Conv}\,3\times 3\left(T_2\right)\right)\right)_{\times 2}. \end{aligned} \tag{1}$$

Subsequently, through the FIM, as depicted in Fig. 3, we process features X_1 and X_2 by performing channelwise concatenation, elementwise multiplication, elementwise absolute difference, and feature maximization operations, which can be articulated by (2), yielding four features X_c, X_m, X_d, and $X_{\max}$, each with dimensions$\in \mathbb{R}^{H\times W\times C}$. These four features are then fused along the channel axis, and a 3×3 convolution operation scales them back to the original number of channels to produce $X_f \in \mathbb{R}^{H\times W\times C}$, as represented by (3)

$$\begin{aligned} X_c &= R\left(B\left(\text{Conv}\,1\times 1\left(\text{Concat}\left(X_1, X_2\right)\right)\right)\right) \\ X_m &= X_1 \times X_2 \\ X_d &= \left|X_1 - X_2\right| \\ X_{\max} &= \text{Max}\left(X_1, X_2\right) \end{aligned} \tag{2}$$

$$X_f = R\left(B\left(\text{Conv}\,3\times 3\left(X_c, X_m, X_d, X_{\max}\right)\right)\right). \tag{3}$$

Next, feature $X_f \in \mathbb{R}^{H\times W\times C}$ is introduced into the DE unit, as shown in Fig. 4, where feature X_f is further refined and enhanced through both local and global attention mechanisms. In the local attention branch, channel compression is achieved by 1×1 convolutional layer, batch normalization and relu activation function are used to improve the nonlinear expressive power, and then the number of channels is recovered by 1×1 convolutional layer to obtain $X_{f1} \in \mathbb{R}^{H\times W\times C}$. In the global attention branch, adaptive average pooling is used to generate global features to capture the overall spatial distribution statistics of the scene, and the same channel compression and recovery is performed to obtain $X_{f2} \in \mathbb{R}^{H\times W\times C}$. The two branches are then each fused by an additive operation with initial features T_1 and T_2. Subsequently, the features $\in \mathbb{R}^{H\times W\times C}$ obtained by fusing the two branches through additive operations are fused with the initial features T_1 and T_2 using the Sigmod function and they use residuals to obtain $X_{f_1}{}' \in \mathbb{R}^{H\times W\times C}$ and $X_{f_2}{}' \in \mathbb{R}^{H\times W\times C}$. Subsequently, the bidimensional temporal features are spliced in the channel dimensions, and the spatial resolution of the feature maps is reduced by using a 3×3 convolution to maintain the spatial resolution of the feature maps to get $X_f{}'$. as expressed in (4). This methodology not only preserves the original feature information but also enhances its representation, thereby improving the detection capability for changes, effectively highlighting key information while suppressing background noise

$$\begin{aligned} X'_{f_1} &= \delta\left(X_{f_1} + X_{f_2}\right) \times T_1 + T_1 \\ X'_{f_2} &= \delta\left(X_{f_1} + X_{f_2}\right) \times T_2 + T_2 \\ X'_f &= R\left(B\left(\text{Conv}3\times 3\left(\text{Concat}\left(X'_{f_1}, X'_{f_2}\right)\right)\right)\right). \end{aligned} \tag{4}$$

Ultimately, feature X_f is fused with the channel-reduced feature through residual connections, further enriching the diversity and information content of the features. The module outputs the

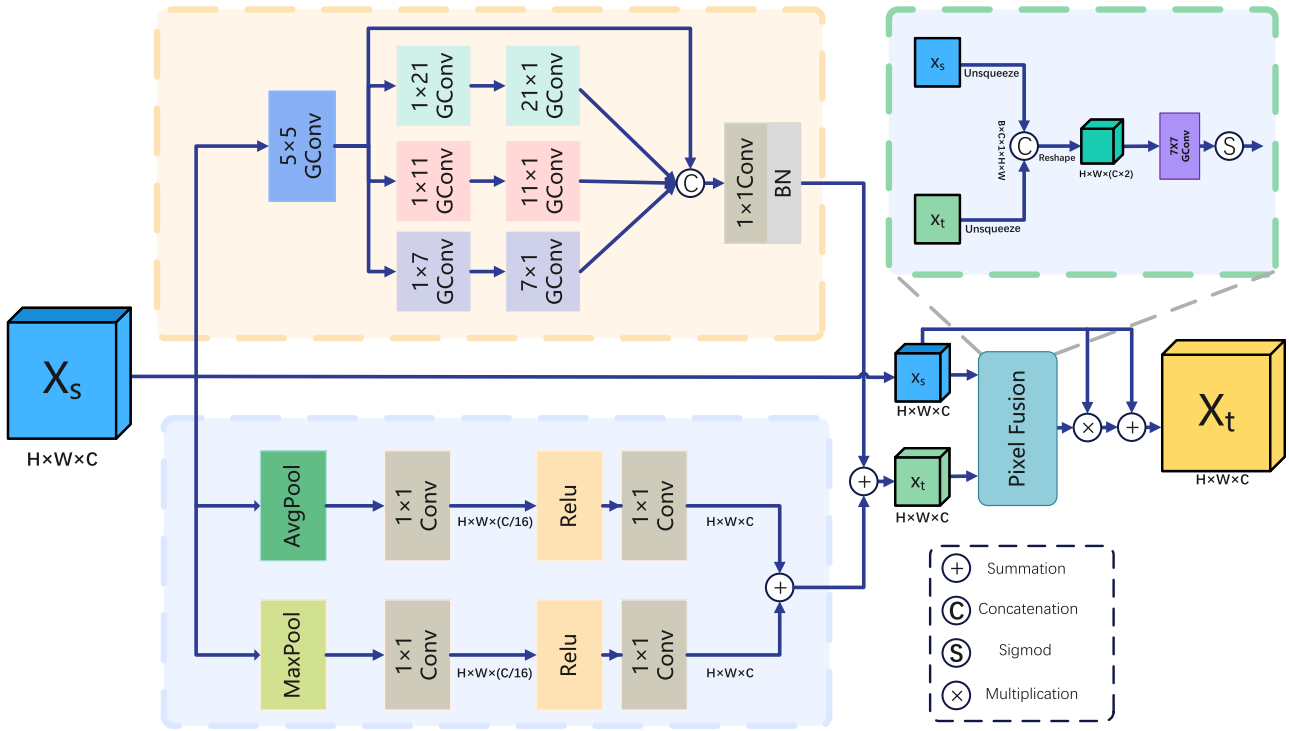

Fig. 5. Depicts the process of the MSGA module, involving pixel-level fusion of features xt and xs obtained through multiscale and channel fusion.

result $X_s \in \mathbb{R}^{H\times W\times C}$ processed by TIDEM through a 1×1 convolution layer.

C. Multiscale Global-Aware Module

Multiscale and global awareness are two commonly employed concepts in computer vision and image processing. Multiscale processing involves considering an image at various scales during its analysis. Objects and environmental features with multiscale characteristics in remote sensing and natural images often contain rich scene information. Global awareness typically describes a characteristic of image processing and computer vision algorithms capable of perceiving and understanding the entire image or global contextual information provided. In remote sensing image analysis, focusing solely on local information may overlook the interconnections between distant objects, failing to accurately comprehend the global semantic information of the image. Hence, we propose MSGA, incorporating convolution kernels of varying receptive fields and fusing feature maps from different scales. Concurrently, a channel attention mechanism is introduced, capturing relationships between channels across the global context and adjusting the feature responses of each channel. By fusing multiscale global feature information with the original feature information at the pixel level, diverse dimensions, including color, texture, and shape, are extracted. The amalgamation of these pieces of information aids in better distinguishing different types of land cover, especially in complex terrestrial environments. In summary, MSGA enables the model to capture key feature information in crucial areas within remote sensing imagery, enhancing the accuracy of CD.

Within MSGA, we designed a parallel structure, as illustrated in Fig. 5, treating the channel attention module and the multiscale module as two concurrent processing streams. Initially, the feature map is fed into the channel attention module, undergoing average pooling and max pooling, followed by passage through two 1×1 convolution layers (equivalent to two fully connected layers) for each channel. As demonstrated in (5), nonlinearity is introduced between the first and second fully connected layers via the relu activation function. After processing through the second fully connected layer, the outputs of the two channel attention groups, X_{s1} and X_{s2}, are summed to produce $l_1 \in \mathbb{R}^{H\times W\times C}$

$$\begin{aligned} X_{s1} &= \text{Conv1} \times 1\left(R\left(\text{Conv}\, 1 \times 1\left(\text{AvgPool}\left(X_s\right)\right)\right)\right) \\ X_{s2} &= \text{Conv1} \times 1\left(R\left(\text{Conv1} \times 1\left(\text{MaxPool}\left(X_s\right)\right)\right)\right) \\ l_1 &= X_{s1} + X_{s2}. \end{aligned} \tag{5}$$

The feature map data is then input into the multiscale module, which defines a series of convolution layers of varying scales. This includes variations in convolution kernel sizes, such as 5×5 grouped convolutions, 1×7 grouped convolutions, and 7×1 grouped convolutions, among others. These convolution operations are independently computed across channels to extract spatial relations for each channel individually. The outcome from the 5×5 convolution layer, along with the results M_1, M_2, M_3 obtained from three different convolution operations, are processed in parallel across the channel dimension, as

indicated by the following:

$$\begin{aligned} M &= G\,\mathrm{Conv}\,5\times 5\,(X_s)\\ M_1 &= GConv7\times 1(GConv1\times 7(M))\\ M_2 &= GConv11\times 1(GConv1\times 11(M))\\ M_3 &= GConv13\times 1(GConv1\times 13(M))\\ l_2 &= B\,(\mathrm{Conv}1\times 1\,(Concat\,(M,M_1,M_2,M_3)). \end{aligned} \tag{6}$$

Subsequently, the information from different channels is combined and normalized through a 1×1 convolution layer and normalization, where the resultant outputs $l_2\in\mathbb{R}^{H\times W\times C}$ and $l_1\in\mathbb{R}^{H\times W\times C}$ are summed to generate the feature fusion $X_t\in\mathbb{R}^{H\times W\times C}$. Features X_t and X_s are initially augmented with a new dimension $X_t'\in\mathbb{R}^{1\times H\times W\times C}$ and $X_s'\in\mathbb{R}^{1\times H\times W\times C}$, and then concatenated along this new dimension, as depicted in (7), forming a feature map stack encompassing two time points. This provides the model with a novel perspective for a more nuanced understanding and analysis of remote sensing data. Here, Re denotes Rearrange, which merges the temporal axis with the channel axis $\in\mathbb{R}^{1\times H\times W\times(C\times 2)}$, offering a tensor for subsequent convolution layers as input

$$\begin{aligned} X_t' &= \mathrm{Unsqueeze}\,(X_t)\\ X_s' &= \mathrm{Unsqueeze}\,(X_s)\\ R &= \mathrm{Re}\,(\mathrm{Concat}\,(X_t',X_s'))\\ l_3 &= GConv7\times 7(R)\\ X_t &= \delta\,(l_3)\times X_s + X_s. \end{aligned} \tag{7}$$

A 7×7 grouped convolution is utilized to process this merged feature map, producing the final attention-weighted feature map $l_3\in\mathbb{R}^{H\times W\times C}$. Each channel is assigned a weight through the δ function, effectively reweighting the features by multiplying the weights with the input features themselves. This captures extensive contextual information to compute attention weights for each pixel, dynamically assessing the importance of each pixel within the image. This approach enhances the model's capability to capture and analyze the nuanced changes in land cover between two time points. Finally, information fusion and residual connections with the outputs of the two modules and the original features not only enhance the focus on significant features but also dynamically adjust feature representations while preserving the original information. This improves the model's capacity to process complex data and enhance predictive performance.

D. DA Module

In the task of RSCD, it is imperative to integrate both low-level features, rich in spatial details providing precise spatial localization information, and high-level features, offering robust classification and recognition capabilities with an understanding of semantic information of objects. Hence, through resolution fusion—merging features of different resolutions–the model can simultaneously leverage spatial granularity and semantic comprehension, thereby enhancing the performance of the task. We have developed a dual-source attention block and spatio-temporal coordinate awareness fusion module, capable of deeply merging decoded features across different levels, emphasizing key information from a high receptive field, and ensuring the prominence of crucial information while suppressing nonessential details. This not only augments the model's focus but also improves the representativeness of features.

In response, we introduce a novel DA module, as depicted in Fig. 6. Within the DA, we incorporate a dual-source attention (DSA) block and a spatio-temporal coordinate awareness fusion module. Initially, we input features of both high and low resolutions into the DSA. The first step involves deconvolving feature $X_h\in\mathbb{R}^{\frac{H}{2}\times\frac{N}{2}\times(C\times 2)}$nd passing feature $X_l\in\mathbb{R}^{H\times W\times C}$ through a 1×1 convolution layer followed by normalization, then summing them and applying the nonlinear activation function to obtain feature C. Subsequently, feature C s processed through a 1×1 convolution layer and δ function, and multiplied by the input X_h' to produce the output $X''\in\mathbb{R}^{H\times W\times C}$ of the DSA, as shown in (8). This weighted mechanism enhances features with higher weight values at corresponding locations while suppressing those with lesser attention values, highlighting key areas of the feature. This approach aids in better preserving the detailed information of the original image, thereby improving task-specific performance

$$\begin{aligned} X_h' &= B\,(\mathrm{DeConv}\,(X_h))\\ C &= R\,(X_h'+X_l)\\ X_h' &= \delta(B(Conv1\times 1(C)))\times X_h'. \end{aligned} \tag{8}$$

Next, features X_l and X_h'' are concatenated along the channel dimension, and a combination of a 1×1 convolution layer, batch normalization, and activation layer yields the fused feature $D\in\mathbb{R}^{H\times W\times C}$, facilitating feature extraction and spatial information processing, as illustrated in (9). The result is then input into the spatio-temporal coordinate awareness fusion module for adaptive average pooling operations with respect to height $X_x\in\mathbb{R}^{H\times W\times 1}$ and width $X_h\in\mathbb{R}^{H\times 1\times W}$. The two feature values are then dimensionally concatenated and sliced, processed through a 1×1 convolution, and δ to derive the attention map's weights. Features X_l and X_h'' are adjusted using attention weights to produce X_{l1}, X_{h1}', and the features are elementwise added and passed through a 1×1 convolution group and D for a residual connection to obtain the final result $X_n\in\mathbb{R}^{H\times W\times 1}$. This preserves the original feature information while incorporating attention-enhanced areas of focus

$$\begin{aligned} D &= R\,(B\,(\mathrm{Conv}\,1\times 1\,(\mathrm{Concat}\,(X_h',X_l))))\\ X_x &= \mathrm{AvgPool}_x(D), X_h = \mathrm{AvgPool}_h(D)\\ X_n &= R\,(B\,(\mathrm{Conv}\,1\times 1\,(\mathrm{Concat}\,(X_{h1}',X_{l1}))))+X_h'. \end{aligned} \tag{9}$$

By paying attention to the features at each coordinate, the decoder learns to balance between varying degrees of detail and semantics, aiding in restoring the original input scale while retaining more useful information. Finally, an upsampling and convolution block, as shown in the following, outputs the final

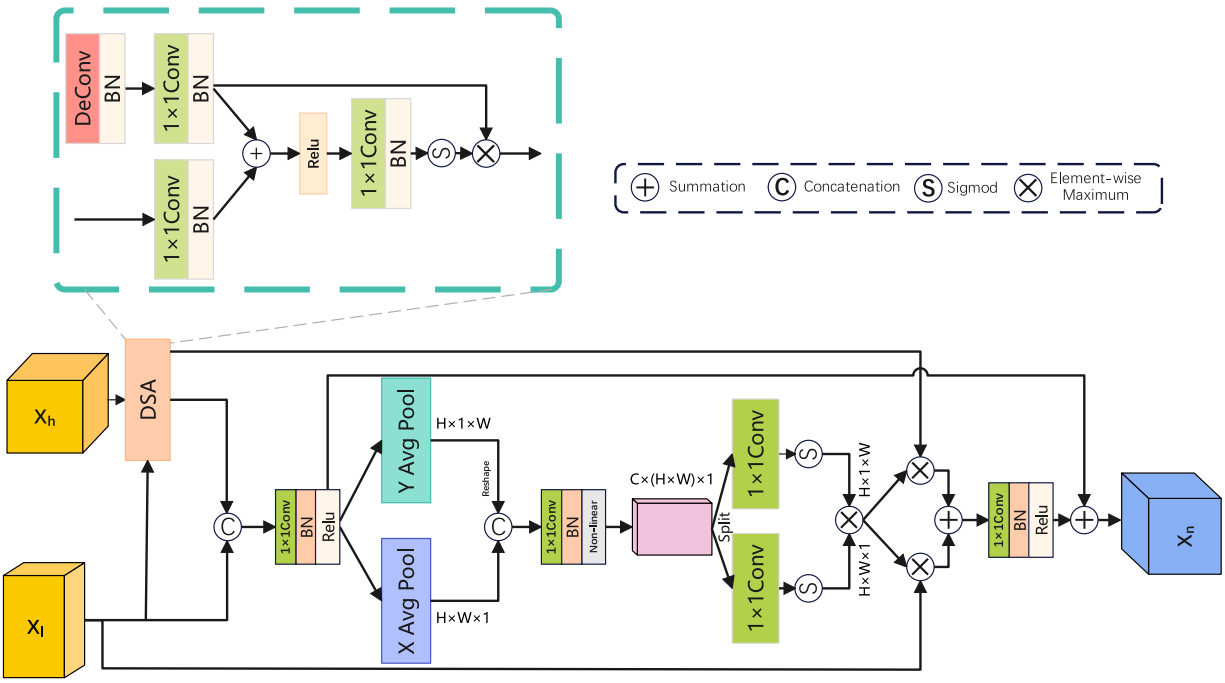

Fig. 6. DA module.

prediction image $X_{\text{out}} \in \mathbb{R}^{H\times W\times 1}$:

$$\begin{aligned} X_n' &= Un\,\text{samplex}\,(X_n) \\ X_n'' &= R\,(B\,(\text{Conv}\,3\times 3\,(X_n'))) \\ X_{\text{out}} &= B\,(\text{Conv}\,3\times 3\,(X_n''))\,. \end{aligned} \tag{10}$$

E. Loss Function

In the field of RSCD, the application of the cross-entropy (CE) loss function offers significant advantages. As a method to measure the discrepancy between the predicted probability distribution and the actual distribution, the CE function precisely calculates the inconsistency between model outputs and true labels, effectively guiding the learning process of the model. Its primary advantage lies in providing continuous and sensitive feedback, ensuring accurate CD in the complex scenarios of remote sensing imagery. Particularly in handling highly imbalanced class distributions, the CE function increases focus on minority classes by penalizing incorrect predictions, driving the model towards optimization and thereby enhancing the overall performance of RSCD tasks. The loss function is defined as shown in the following:

$$L = \frac{1}{H_0\times W_0}\sum_{h=1,w=1}^{H,W} l\,(P_{hw}, Y_{hw}) \tag{11}$$

where $l(P_{\text{hw}}, y) = -\log(P_{\text{hwy}})$ denotes the CE loss, P_{hw} and Y_{hw} represent the labeled and predicted pixel values, respectively. In our network architecture, we employ a deep supervision multiscale loss function strategy, optimizing model performance through the introduction of three different scale auxiliary losses. Based on historical experience, the weights of these auxiliary losses have been meticulously adjusted to 0.5 and 0.2, the aim is to achieve model optimization by further distributing the different losses of the deeply supervised output with different weight values, respectively. Consequently, the comprehensive loss function L of the model can be described as illustrated in (12). The formulas used here for $\text{Loss}_{1,2,3}$ are all computational procedures shown in (11)

$$L = \text{Loss}_1 + 0.5\,\text{Loss}_2 + 0.2\,\text{Loss}_3\,. \tag{12}$$

IV. Experiments

A. Datasets

To confirm the superior performance of TIMF-Net, we conducted tests on three bitemporal RSCD datasets.

The LEVIR-CD [52] dataset, a new large-scale remote sensing building CD dataset, comprises 637 pairs of high-resolution (0.5 m/pixel) Google Earth image pairs, each measuring 1024 × 1024 pixels, with a temporal span of 5–14 years documenting significant land use changes, especially in building growth. This dataset covers various types of buildings, including villas, high-rise apartments, small garages, and large warehouses, focusing on changes related to buildings, both growth and decay. LEVIR-CD contains 31 333 independent building change instances, aiming to provide a new benchmark for evaluating CD algorithms, particularly those based on deep learning. The 637 image pairs were processed and randomly cropped to obtain nonoverlapping images with a resolution of 256 × 256 pixels. These images were then allocated to the test, training, and validation sets in a 2:7:1 ratio.

The WHU-CD [53] dataset, developed by Wuhan University, aims to advance CD technology in the remote sensing

field, with a particular emphasis on surface changes during urbanization processes. It meticulously records the urban architectural changes following the magnitude 6.3 earthquake in Christchurch, New Zealand, in 2011. By including high-resolution remote sensing image pairs before and after the earthquake, WHU-CD offers firsthand visual material for researchers to analyze and understand the impact of natural disasters on urban infrastructure. Large image pairs were segmented into 256×256 pixel blocks, randomly divided into 6096 for training, 762 for validation, and 762 for testing.

The GZ-CD [54] dataset is a high-resolution satellite imagery CD dataset with a spatial resolution of 0.55 m/pixel, covering images of the outskirts of Guangzhou, China, from 2006 to 2019. A total of 19 pairs of images of seasonal changes, with resolutions ranging from 1006×1168 to 4936×5224 pixels, reflecting seasonal variations. These images were collected through the Google Earth service. We also uniformly segmented the images into 256×256 sizes. These blocks were randomly divided into training, validation, and test sets: 2834, 400, and 325, respectively.

B. Implementation Details and Evaluation Metrics

1) Implementation Details: The model was built using Python 3.7 and PyTorch 1.13.1, trained on a machine with 12 G memory NVIDIA RTX 2080Ti. The network batch size was 10. The CE loss function was employed as the loss function. AdamW was used as the optimizer, with an initial learning rate of 0.001, weight decay of 0.01, and beta values of (0.9, 0.999).

2) Evaluation Metrics: To facilitate comparison with other advanced methods, we selected five evaluation metrics: F1, IoU, Precision, Recall, OA. Among these, F1 score and IoU are the primary metrics, defined as follows:

$$
\begin{aligned}
F1 &= 2\frac{\text{Precision-Recall}}{\text{Precision} + \text{Recall}} \\
\text{Precision} &= \frac{TP}{TP + FP} \\
\text{Recall} &= \frac{TP}{TP + FN} \\
\text{IoU} &= \frac{TP}{TP + FN + FP} \\
\text{OA} &= \frac{TP + PN}{TP + FN + FP}.
\end{aligned} \tag{13}
$$

C. Contrast Experiment

In the contemporary research landscape, particularly with the ongoing advancements in RSCD technology, the adoption of cutting-edge deep learning strategies has become a pivotal pathway for significantly enhancing detection precision and efficiency. To demonstrate the superior performance and high accuracy of our model—TIMF-Net—in RSCD tasks, we meticulously compared it against several models that have achieved breakthrough accomplishments in this field. We conducted comparative experiments using the same general parameters and environment, ensuring an identical number of epochs. These models include three dual-stream network structures based on classical convolution: early fusion convolutional encoder (FC-EF), fully convolutional Siamese networks for difference (FC-Siam-Di), and fully convolutional Siamese networks with concatenation (FC-Siam-Conc); three advanced methods that incorporate deep convolutional network attention mechanisms: IFNet, Siamese nested U-Net (SNUNet), DMINet; two innovative approaches using Transformer technology: BIT and ChangeFormer; along with two unique models that merge the advantages of CNNs and Transformers: feature transformation network (FTN) and ICIF-Net. The following text briefly outlines the core concepts of these nine models, revealing their contributions and distinguishing features in the realm of RSCD.

FC-EF [55] is a convolutional encoder structure designed for early fusion, specifically tailored to process paired input images. By fusing the features of paired images at an early stage in the input layer, FC-EF can effectively capture and utilize the correlational information between these images, thereby enhancing processing efficiency and model performance.

FC-Siam-Di [55] is a dual-stream network structure based on a fully convolutional network, aimed at detecting changes by computing differences between paired remote sensing images. This model deeply extracts features from images of two different times, then processes these features to identify changed areas between the images.

FC-Siam-Conc [55] is a fully convolutional dual-stream network designed for RSCD. Unlike other versions, FC-Siam-Conc employs a concatenation approach to process paired remote sensing images after feature extraction.

IFNet [56] is a deep learning framework designed for feature fusion and analysis of images, especially in fields like remote sensing image processing and CD. It captures rich spatial information through multiscale convolutional kernels, achieving deep integration of image features.

SNUNet [57] is a deep learning architecture specifically designed for complex image segmentation and CD tasks, especially in the analysis of remote sensing images. Based on the classic U-Net architecture and employing nesting techniques, it effectively enhances the capability to capture features of remote sensing images and the precision in identifying change areas.

FTN [58] leverages the Swin Transformer as a core feature extraction tool, effectively capturing rich multiscale feature information. By combining a pyramid structure with progressive attention modules, FTN accurately processes and analyzes multilevel information extracted from the backbone network, focusing meticulously on changed targets.

BIT [30] is a Transformer model for processing bitemporal images, which translates images into semantic tokens to effectively capture spatio-temporal context. Utilizing the encoder–decoder of Transformer, BIT significantly enhances CD performance with minimal computational resources, even based on a simple ResNet18 structure.

ChangeFormer [29] is a dual-stream network based on Transformer technology for remote sensing image CD, demonstrating superior performance by integrating multiscale details, surpassing traditional fully convolutional network frameworks.

TABLE I
RESULTS OF INDICATORS ON THE LEVIR-CD DATASET FOR EACH COMPARISON METHOD

Method	LEVIR-CD				
	Precision	**Recall**	**F1**	**IoU**	**OA**
FC-EF	86.91	80.17	83.40	71.53	98.39
FC-Siam-Di	89.53	83.31	86.31	75.92	98.67
FC-Siam-Conc	91.99	76.77	83.69	71.96	98.49
IFNet	94.02	82.93	88.13	78.77	98.87
SNUNet	89.18	87.17	88.16	78.83	98.82
BIT	89.24	89.37	89.31	80.68	98.92
ChangeFormer	92.05	88.80	90.40	82.48	99.04
ICIF-Net	91.13	90.57	91.18	83.85	99.12
FTN	92.71	89.37	91.01	83.51	99.06
DMINet	92.52	89.95	90.71	82.99	99.07
AERNet	89.97	91.59	90.78	83.11	99.07
SRCNet	93.50	90.07	91.75	84.76	99.17
TIMF-Net	92.78	91.15	91.96	85.12	99.18

Red represents the best indicator value, blue represents the second best indicator value.

ICIF-Net [51] combines the advantages of CNNs and Transformers through a Conv Attention module to facilitate interaction between local and global features and employs attention mechanisms for cross-scale fusion, effectively integrating multiresolution information. Ultimately, it outputs precise results through a change prediction head, showcasing a new perspective on combining CNNs and Transformers.

DMINet [59] combines self-attention and cross-attention in a single module and introduces a joint-attention (JoinAtt) module across time and uses subtraction and join operations to achieve aggregation of multilevel feature differences. This is designed to improve processing efficiency and optimize the feature integration process.

AERNet [60]: This network enhances feature extraction of changes in building structures by using an attentional mechanism to refine edge details. This approach helps to identify and analyze building changes more accurately, especially in complex remote sensing images.

SRCNet [61]: This network for CD in remote sensing images, aiming to fully utilize the spatial relationships in dual-temporal images. It contains two key modules: the perception and interaction module, which improves the accuracy and robustness of feature extraction through the cross-branch perception mechanism.

On the LEVIR-CD dataset, as illustrated in Table I, we compared TIMF-Net against other exemplary models, finding it significantly outperforms the existing advanced models across multiple crucial performance metrics. It is particularly noteworthy that our model exhibits superior Recall, F1, IoU, and OA scores compared to others, with the most critical F1 and IoU metrics standing at 91.96% and 85.12%, respectively, markedly surpassing the second-ranked ICIF-Net by increments of 0.78% and 1.27%. Furthermore, it also achieves higher Recall and OA rates than ICIF-Net by 0.58% and 0.06%, respectively. Overall, our model demonstrates superiority over other advanced algorithms on the LEVIR-CD dataset, a success attributable to our MSGA module, which offers significant advantages in both local and global processing across various scales, particularly in edge analysis and the detection of large buildings. The data in Table I corroborates the feasibility, effectiveness, and superiority of TIMF-Net. Among models utilizing pure Transformer or ResNet as their backbone network, selecting PVTb2 as the backbone has yielded conspicuous results. Fig. 7 provides an intuitive comparison of each model, presenting a visual representation of the differences between generated images and the ground truth (GT) across different settings. This not only showcases visible changes along the edges of large buildings but also whether the model can detect small buildings that are difficult to recognize. We utilized four colors for visual representation in the images: black for TN, white for TP, green for FN, and red for FP, where TN represents true negatives, TP true positives, FN false negatives, and FP false positives. Six images depicting buildings of various architectures and styles were selected for comparison. In Fig. 7(a)–(f), our model notably outperforms others, with the smallest proportion of red and green areas and the highest similarity to the actual GT images. Fig. 7(a) and (b) displays two sets of images featuring buildings with closely spaced, intricately detailed edges, where our method distinctly excels in accurately identifying small gaps between buildings while maintaining the integrity of their edges, unlike the nine comparative methods that exhibit various edge detection errors or fail to accurately detect narrow areas, resulting in a significantly larger proportion of green and red areas. Fig. 7(c) and (d) features two sets of images after large building changes, demonstrating our method's ability not only to recognize the overall appearance and complete features of large buildings but also to accurately identify some small buildings outside the large building areas, a feat other methods fail to achieve, lacking completeness and often missing or incorrectly identifying small buildings. Fig. 7(e) and (f) showcases two sets of small building scenarios, where our method's superiority in the integrity and accuracy of small targets is evident compared to other models. These comparative images unequivocally prove our model's outstanding global and local advantages on the LEVIR-CD dataset over other models.

On the WHU dataset, as demonstrated by the results presented in Table II, TIMF-Net achieves an F1 score of 93.37%, an IoU of 87.56%, an OA of 99.38%, and a Recall of 91.45%, all surpassing the metrics of other models. Compared to the second-ranked FTN model, TIMF-Net's performance is superior, with a 1.21% higher F1 score, 2.11% higher IoU, 0.21% higher Recall, and 0.01% higher OA. This exceptional performance highlights TIMF-Net's distinct superiority in processing the WHU dataset. Upon in-depth analysis, we attribute its success to the effective integration of multitemporal feature information by TIMF-Net and its innovative feature fusion technology that precisely captures the details of change areas. TIMF-Net's success in showcasing the impact of natural disasters on urban infrastructure, especially in complex urban environments and diverse land cover types, lies in its ability to accurately identify minor changes and effectively suppress irrelevant information.

Visual comparisons in Fig. 8 clearly reveal TIMF-Net's significant progress on the WHU dataset over other models. Particularly in Fig. 8(a), TIMF-Net successfully avoids misjudgments caused by pseudochanges, capturing target changes

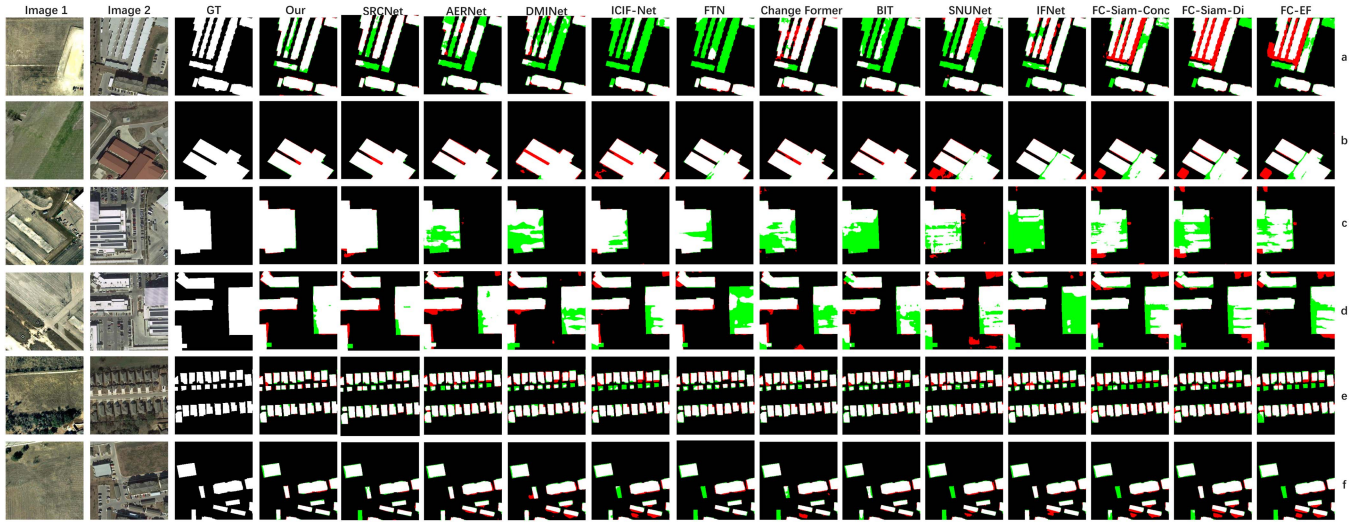

Fig. 7. We compare TIMF-Net with other models in the LEVIR-CD dataset for difference, and use colors such as red and green to interpret the changes, so as to more intuitively reflect the differences between images of different models and real images, in which white represents TPs, black represents TNs, red represents FPs, green represents FNs. We chose (a)–(f) pictures to show the visual comparison results of different images in different models.

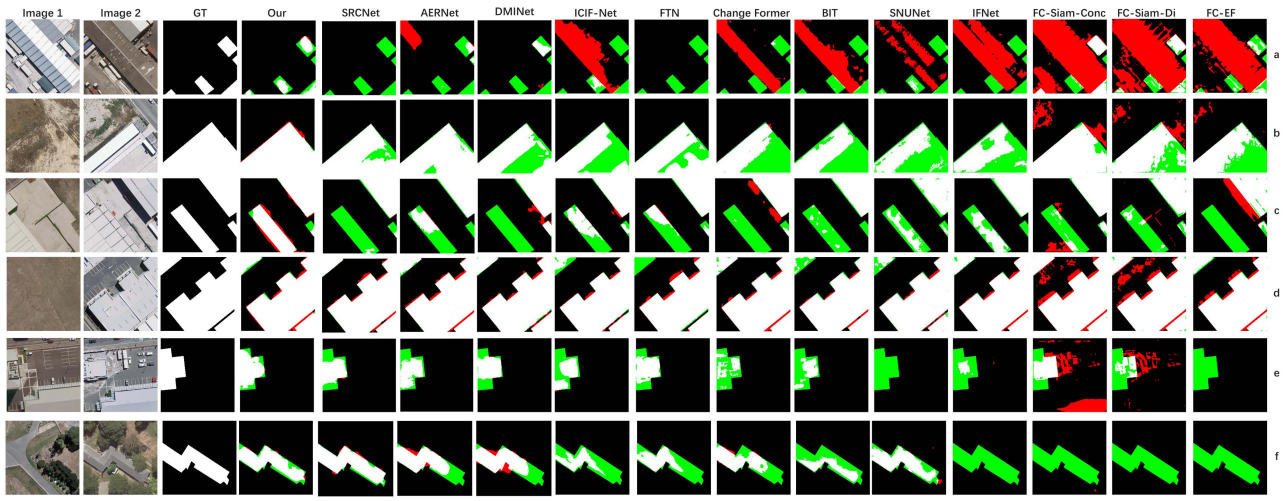

Fig. 8. We compare TIMF-Net with other models in the WHU-CD dataset for difference, and use colors such as red and green to interpret the changes, so as to more intuitively reflect the differences between images of different models and real images, in which white represents TPs, black represents TNs, red represents FPs, green represents FNs. We chose (a)–(f) pictures to show the visual comparison results of different images in different models.

more completely and accurately, whereas other models often miss key change areas or erroneously mark changes in roof colors as areas of change. In the cases of large building changes displayed from Fig. 8(b)–(e), TIMF-Net significantly surpasses competing models in maintaining target integrity and accuracy, especially evident in its ability to finely recognize large-area change targets. Similarly, in scenarios involving small targets, as shown in Fig. 8(e) and (f), TIMF-Net also significantly outperforms other methods in edge detection and maintaining target integrity,validating its efficiency in detail processing and small target recognition. These visual results not only highlight TIMF-Net's technical advantages but further prove its application potential and practical value in the field of RSCD.

On the GZ-CD dataset, the results presented in Table III showcase the exceptional achievements of our model in core performance metrics, notably outperforming the closely following FTN method by a margin of 1.54% in F1 score and 2.4% in IoU, while also maintaining a leading position in precision, recall, and oa, with only slight differences compared to the FTN method. Despite minor shortcomings in individual metrics, our model demonstrates the most outstanding performance both theoretically and practically, showcasing superior results on the GZ-CD dataset. Through a detailed comparative analysis from Fig. 9(a)–(f), we selected six groups of contrast images with significant changes in lighting and seasons. Our model consistently outperforms other methods across various scenarios. In the experiments with significant lighting changes shown in

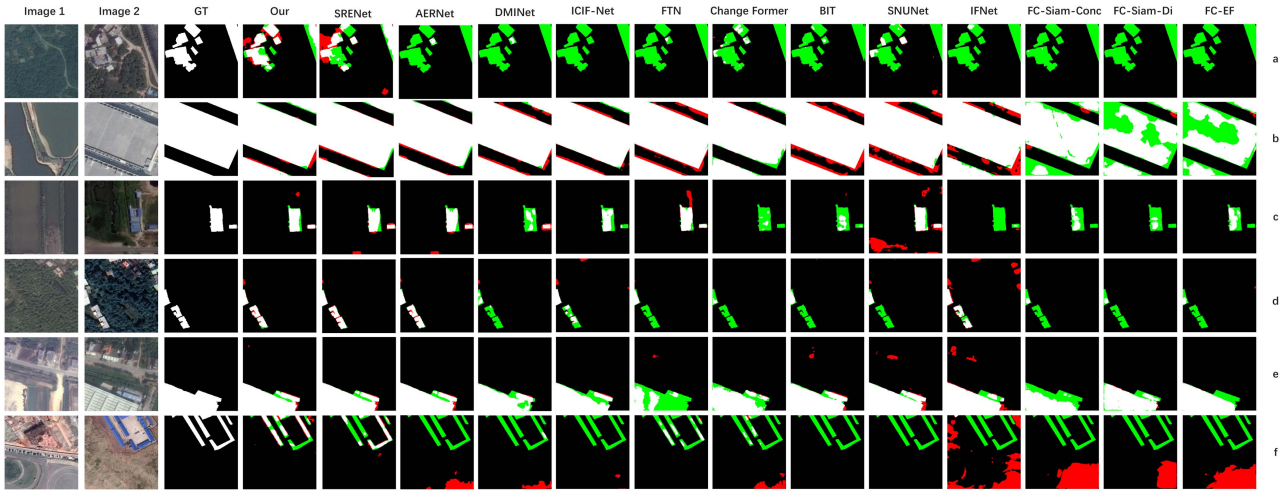

Fig. 9. We compare TIMF-Net with other models in the GZ-CD dataset for difference, and use colors such as red and green to interpret the changes, so as to more intuitively reflect the differences between images of different models and real images, in which white represents TPs, black represents TNs, red represents FPs, green represents FNs. We chose (a)–(f) pictures to show the visual comparison results of different images in different models.

TABLE II
RESULTS OF INDICATORS ON THE WHU-CD DATASET FOR EACH COMPARISON METHOD

Method	WHU-CD				
	Precision	**Recall**	**F1**	**IoU**	**OA**
FC-EF	71.63	67.25	69.37	53.11	97.61
FC-Siam-Di	47.33	77.66	58.81	41.66	95.63
FC-Siam-Conc	91.99	76.77	83.69	71.96	98.49
IFNet	96.91	73.19	83.40	71.52	98.83
SNUNet	85.60	81.49	83.50	71.67	98.71
BIT	86.64	81.48	83.98	72.39	98.92
ChangeFormer	90.50	79.61	84.51	73.18	98.59
ICIF-Net	92.93	88.70	90.77	83.09	99.13
FTN	93.09	91.24	92.16	85.45	99.37
DMINet	92.65	90.35	91.49	84.31	99.19
AERNet	92.47	91.89	92.18	85.49	99.25
SRCNet	93.42	92.64	93.03	86.96	99.33
TIMF-Net	95.37	91.45	93.37	87.56	99.38

Red represents the best indicator value, blue represents the second best indicator value.

TABLE III
RESULTS OF INDICATORS ON THE GZ-CD DATASET FOR EACH COMPARISON METHOD

Method	GZ-CD				
	Precision	**Recall**	**F1**	**IoU**	**OA**
FC-EF	71.63	67.25	69.37	53.11	97.61
FC-Siam-Di	47.33	77.66	58.81	41.66	95.63
FC-Siam-Conc	80.59	72.34	76.24	61.61	95.68
IFNet	92.19	74.08	82.15	69.71	96.92
SNUNet	84.25	81.82	84.25	72.79	97.07
BIT	82.40	78.18	80.23	66.99	96.31
ChangeFormer	84.59	65.23	73.66	58.30	95.53
ICIF-Net	89.90	80.76	85.09	74.05	97.29
FTN	86.99	84.21	85.58	74.79	97.92
DMINet	87.92	76.79	81.98	69.46	96.77
AERNet	88.06	81.07	84.42	73.03	97.13
SRCNet	89.35	84.43	86.82	76.71	97.62
TIMF-Net	90.41	84.07	87.12	77.10	97.62

Red represents the best indicator value, blue represents the second best indicator value.

Fig. 9(a) and (d), other methods failed to detect the changes in targets, whereas our approach, despite some detection blur, was still able to identify changes in building clusters. In Fig. 9(b), our model demonstrated precision and finesse in edge detection of large targets undergoing significant changes, highlighting its mastery over complex scenes. In the detection of small buildings as displayed in Fig. 9(c), our technology overcame the effects of seasonal changes and made progress in handling detailed edge processing. Even in the detection of slender target changes shown in Fig. 9(f), while our model exhibited some limitations, it still closely matched the actual changes better than other methods, showcasing its exceptional capability in capturing subtle variations.

Through experiments across three datasets, we have evidenced the superiority of TIMF-Net over other methods, showcasing its distinct advantages in the field of RSCD. Detailed comparative images in Fig. 10 further demonstrate our method's superiority in handling both fine targets and large-area target edges, robustly supporting TIMF-Net's status as an efficient and reliable CD tool, capable of delivering accurate and consistent results in a variety of complex environments. The outstanding performance of our method across different datasets can be attributed to its innovative temporal interaction and difference enhancement techniques, especially its ability to effectively integrate bitemporal features and suppress irrelevant information for refined processing of remote sensing images. Despite TIMF-Net's clear advantages over other methods, there remains room for improvement, such as in handling blurred edges and complex structures as shown in Fig. 9(a) and (f), indicating areas for further efforts.

Computational and parameter counts: In the summary in Table IV , we compare in detail the parameter counts, computational costs, and F1 scores and IoU metrics of the various approaches. Our TIMF-Net model achieves a balanced performance among many competing schemes with a parameter count of 27.64 M and a computational volume of 42.8 G with a

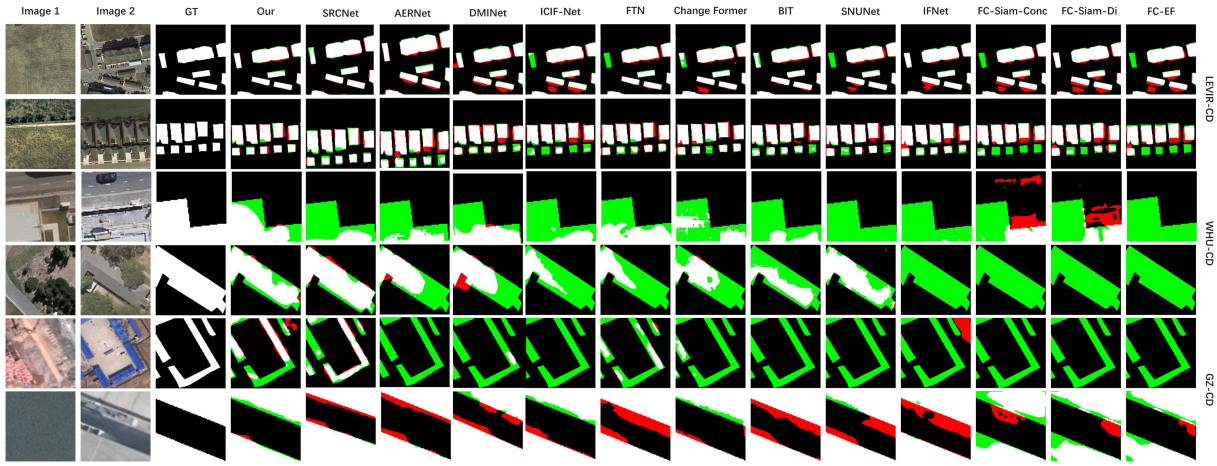

Fig. 10. We made a difference comparison between the more detailed images in the comparison images in the three datasets, and we could clearly observe the changes in the edge region. The images of LEVIR-CD are from Fig. 6(e), the images of WHU-CD are from Fig. 7(e), and the images of (f) GZ-CD are from Fig. 8(c) and (f).

TABLE IV
COMPLEXITY AND NUMBER OF PARAMETERS WITH THE VALUES OF F1 AND IOU ON THE THREE DATASETS

Method	Complexity			LEVIR-CD		WHU-CD		GZ-CD	
	Param	**FLOPs**	**Tt(s)**	**F1**	**IoU**	**F1**	**IoU**	**F1**	**IoU**
FC-EF	1.35	3.57	34.80	83.40	71.53	69.37	53.11	69.37	53.11
FC-Siam-Di	1.35	4.72	39.21	86.31	75.92	58.81	41.66	58.81	41.66
FC-Siam-Conc	1.55	5.32	39.73	83.69	71.96	83.69	71.96	76.24	61.61
IFNet	50.71	82.35	121.24	88.13	78.77	83.40	71.52	82.15	69.71
SNUNet	12.03	54.88	405.3	88.167	78.83	83.50	71.67	84.25	72.79
BIT	3.55	10.59	147.54	89.31	80.68	83.98	72.39	80.23	66.99
ChangeFormer	41.03	202.83	373.43	90.40	82.48	84.51	73.18	73.66	58.30
ICIF-Net	25.83	25.27	270.5	91.18	83.85	90.77	83.09	85.09	74.05
FTN	168.53	45.00	421.67	91.01	83.51	92.16	85.45	85.58	74.79
DIMNet	6.24	14.55	77.22	90.71	82.99	91.49	84.31	81.98	69.46
AERNet	25.36	12.82	156.51	90.78	83.11	92.18	85.49	84.42	73.03
SRCNet	5.17	95.22	191.23	91.75	84.76	93.03	86.96	86.82	76.71
TIMF-Net	27.64	42.80	296.3	91.96	85.12	93.37	87.56	87.12	77.19

TABLE V
ABLATION EXPERIMENTS WERE PERFORMED ON TIDEM; TIDEM'S FIM AND DE WERE ABLATED, AND F1 AND IOU WERE SHOWED ON THE THREE DATASETS

No.	FIM	DE	LEVIR-CD		WHU-CD		GZ-CD	
			F1	**IoU**	**F1**	**IoU**	**F1**	**IoU**
1	×	✓	91.78	84.82	93.13	87.26	86.76	76.56
2	✓	×	91.81	84.90	93.21	87.33	86.88	76.68

training time of 296.3 s. Despite maintaining a moderate level of resource consumption, our approach significantly outperforms other approaches in terms of effectiveness, especially in F1 score and IoU metrics far outperforming more resource-consuming approaches such as FTN, ChangeFormer, and so on.

Noise learning: In order to further verify that the method has better anti-interference and robustness, we use Gaussian white noise with 10 db interval for 10–30 db variable addition before image preprocessing, and test it on all the comparison experiments and TIMF-Net, as shown in Fig. 11, which can be intuitively observed that the F1 scores of our method in different phases of the noise are better than the other methods, reflecting

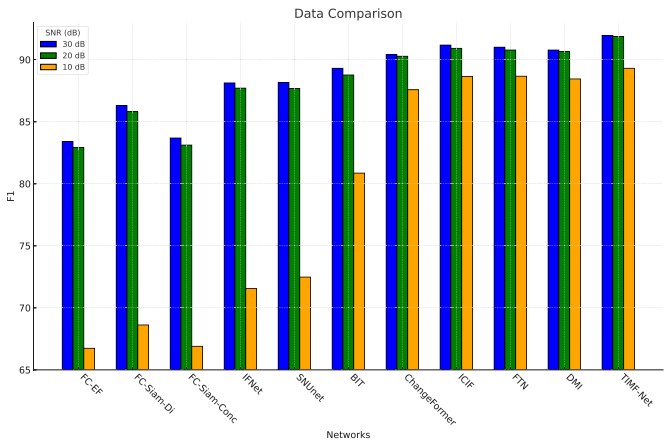

Fig. 11. Different Gaussian noise.

TABLE VI
ABLATION EXPERIMENTS OF OUR METHOD WERE PERFORMED ON THREE DATASETS; THE TIDEM, MSGA, AND DA MODULES WERE ABLATED AND THE VALUES OF F1 AND IOU WERE CALCULATED

No.	TIDEM	MSGA	DA	LEVIR-CD		WHU-CD		GZ-CD	
				F1	**IoU**	**F1**	**IoU**	**F1**	**IoU**
1	×	✓	✓	91.57	84.45	92.18	85.49	84.41	73.04
2	✓	×	✓	91.62	84.54	92.31	85.93	85.76	75.07
3	✓	✓	×	91.70	84.71	92.82	86.96	86.12	75.61
4	✓	×	×	89.22	79.36	88.01	78.42	74.81	61.88
5	×	✓	×	88.34	79.17	86.71	76.54	72.20	56.49
6	×	×	✓	86.74	76.59	85.32	74.82	68.75	51.11
7	×	×	×	85.47	74.82	78.35	64.40	66.71	50.04

the fact that TIMF-Net has the best performance and highest robustness against noise effects.

D. Ablation Experiment

To validate the indispensability of each module in our method, we conducted ablation experiments on the TIDEM, MSGA, and DA modules, demonstrating the efficacy and innovation of our approach. The results, as shown in Table VI, provide a detailed evaluation.

In the modification of the TIDEM module, we initially replaced the FIM module with a basic channel fusion strategy, employing 1×1 convolutional kernels for channel dimension adjustment. This ensured that while simplifying the model's structure, the original channel scale was maintained. The purpose of this step was to remove the original complex fusion interaction module, simplifying it to a basic fusion process of bitemporal feature maps, and concurrently omitting elementwise multiplication, elementwise absolute difference calculations, and feature maximum value selection operations. Moreover, the DE module was removed, indicating that when processing fused features, the model would no longer distinguish between global and local information. To ensure fairness in the ablation experiments, weights were assigned to the fused features through the Sigmoid function, maintaining consistency with the original weighting method. Results shown in Table V independently evaluated the ablation effects of the FIM and DE modules. Removal of the FIM module alone, due to the lack of diversified feature fusion, resulted in a decrease in performance and accuracy in surface CD across three datasets. Specifically, the F1 scores decreased by 0.18%/0.24%/0.36%, while the IoU metrics decreased by 0.3%/0.3%/0.63%, respectively. When the DE module was removed independently, the absence of global and local information in the fused features allowed some noise to emerge, leading to a performance decline across the datasets. Specifically, F1 scores decreased by 0.15%/0.16%/0.46%, while IoU decreased by 0.22%/0.44%/ 0.51%, respectively. For the comprehensive ablation of the TIDEM module, both modules were simultaneously removed to fully assess the contribution of the TIDEM module to the overall detection performance of the model. By executing similar steps as in the single-module ablation, we aimed to meticulously explore the role and impact of the TIDEM module as a whole within the model. According to data presented in Table VI, the absence of the TIDEM module's processing led to the model's inability to effectively identify complete target changes, causing a significant discrepancy between the generated images and those processed by our method. The F1 scores decreased by 0.39%/1.19%/2.71%, and IoU decreased by 0.67%/2.07%/4.15%. This difference is attributed not only to the reduction in parameters but more importantly to a severe decline in the model's ability to capture key change information, when the TIDEM module was added separately to the baseline model, F1 increased by 3.75%/9.66%/8.1% and IoU increased by 4.54%/14.01%/11.84%. Therefore, the absence of the TIDEM module directly impacts the model's ability to interpret complex change scenarios, resulting in a dual loss of precision and robustness in CD accuracy in practical applications, further confirming the importance of the TIDEM module in enhancing the overall performance of the model.

In the ablation experiments on the MSGA module, we removed the global and multiscale modules and the pixel attention module.The MSGA module is unique in its ability to integrate feature maps at different scales, capture both the global field of view and the detail level information, and effectively protect the edge information from being lost through a fine pixel-level extraction mechanism. This integrated processing strategy ensures that the model is able to balance overall coherence with local detail accuracy when processing complex images. The removal of the MSGA module had a significant negative impact on the model performance. Specifically, on the three different datasets, the F1 score went down by 0.34%/0.55%/1%, while the IoU metrics were even lower by 0.41%/0.6%/1.58%. In the benchmark model No.7, the F1 score increased by 2.87%/8.36%/5.49%, while the IoU metrics decreased by 4.35%/12.14%/6.45%. These data not only show reveal that the model's ability in global consistency, multiscale fusion processing, and detail capture is significantly weakened in the absence of the MSGA module. In MSGA, in order to demonstrate that pixel attention can better capture the nuances and important information in the image, we removed the pixel attention module, as shown in No. 2 of Table VII, and it can be clearly observed that the F1 score decreased by 0.12%/0.24%/0.82% and the IoU decreased by

TABLE VII
ABLATION EXPERIMENTS OF OUR METHOD WERE PERFORMED ON THREE DATASETS; THE MSGA AND PIXEL FUSION MODULES WERE ABLATED AND THE VALUES OF F1 AND IOU WERE CALCULATED

No.	Module	LEVIR-CD		WHU-CD		GZ-CD	
		F1	**IoU**	**F1**	**IoU**	**F1**	**IoU**
1	TIMF-Net	91.96	85.12	93.37	87.56	87.12	77.19
2	Pixel	91.84	84.91	93.25	87.36	86.30	76.43
3	ASPP	91.73	84.73	92.67	86.35	86.01	75.48
4	FPN	91.66	84.60	92.53	86.10	85.95	75.36

TABLE VIII
ABLATION EXPERIMENTS WERE PERFORMED ON TIDEM; TIDEM'S DSA AND SCF WERE ABLATED, AND F1 AND IOU WERE SHOWED ON THE THREE DATASETS

No.	DSA	SCF	LEVIR-CD		WHU-CD		GZ-CD	
			F1	**IoU**	**F1**	**IoU**	**F1**	**IoU**
1	×	√	91.84	84.91	93.21	87.36	86.92	76.96
2	√	×	91.87	84.97	93.01	87.29	86.85	76.88

0.21%/0.2%/0.76%. Besides, in No.3 and No.4 of Table VII, in order to prove that MSGA is superior to other multiscale modules, two types of traditional multiscale modules, ASPP [62] and FPN [63], are selected for comparison, and according to the experimental results, a single traditional multiscale module lacks sufficient mechanisms to integrate these features effectively, and shows a drop in both F1 and IoU compared to MSGA. The MSGA module not only improves the performance index of the model, but also provides the model with the depth ability to understand and process complex images, which is of great significance to enhance the practical application value of the model.

In the ablation experiment on DA module, we removed the DSA block and spatio-temporal coordinate-aware fusion module, which fuses the features of different layers at the same time coordinate information integration processing, cohesion of the main features of different layers to prevent the loss of key features. After removing the modified module, according to Table VI, we can clearly observe that the parameters of the dataset have been reduced, in which the F1 scores have been reduced by 0.26%/0.39%/0.76%,and the IoU by 0.41%/0.6%/1.58%. In Benchmark Model No.7, with the addition of the DA module, F1 is up 1.26%/6.96%/2.04%, IoU is up 1.77%/10.41%/1.07%. In order to further demonstrate the advantages of the modules, we conducted separate ablation experiments for the DSA module and the spatio-temporal coordinate perception fusion module, as shown in Table VIII No.1, in the DSA ablation, we chose to remove this module, the DSA integrates the information from the high and low sensory fields to emphasize the most important features in the scale. We can see the F1 score was reduced by 0.12%/0.16%/0.2% and the IoU was reduced by 0.21%/0.2%/0.23%. In the spatio-temporal coordinate-aware fusion SCF module ablation experiments, as shown in Table VIII No.2, we also removed the entire module, which focuses on features at each coordinate that can be learned

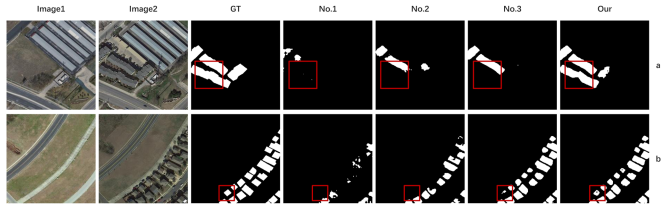

Fig. 12. LEVIR-CD dataset's ablation study outcomes.

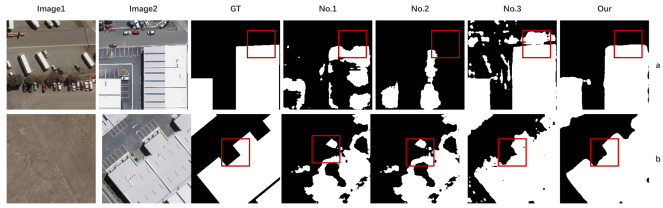

Fig. 13. WHU-CD dataset's ablation study outcomes.

to tradeoff between varying levels of detail and semantics, and the F1 score reduced by 0.09%/0.15%/0.2%0.27% and the IoU by 0.15%/0.27%/0.31% after removing the module. These data not only visualize the exact magnitude of the decrease in performance metrics, but more importantly, they reflect the model's significantly diminished ability to capture and maintain critical feature information after the loss of the DA module.

In our ablation experiments across three datasets, No.1 shows results without the TIDEM module, No.2 without MSGA, No.3 without DA, and our reflects the full model's results. Fig. 12 provides visual results on the LEVIR-CD dataset, showing the impact of different ablation experiments in specific areas (marked with red boxes). These visualizations clearly show that each ablation experiment introduced specific issues. For instance, the results of Experiment No.1 demonstrated a complete failure to detect targets in the first set of samples, while only a few targets were detected in the second set. Although Experiment No.2 detected some targets in the first set of samples, there were missing integrity and edge information, with small targets undetected in the second set. The results of Experiment No.3, similar to No.2, showed slightly finer edge detection but still failed to capture building changes within the red box areas. The detection results for the second set of samples showed some improvement over other ablation experiments but still lacked in detecting small targets. Fig. 13 presents the ablation experiment visual results on the WHU-CD dataset, with two sets of samples emphasizing performance in detecting large-area targets. Results indicated that experiments No.1 and No.2 encountered significant information loss in detecting large-area targets and capturing edge information in the first set of samples, leading to numerous missed detections. In the second set of samples, both experiments also displayed similar issues of information loss. Compared to the first two, experiment No.3 showed slight improvement, managing to vaguely identify building outlines, though edge detection performance was not particularly outstanding, with some false detections in the first set of samples. Fig. 14 displays the visual outcomes of ablation experiments conducted on the GZ-CD

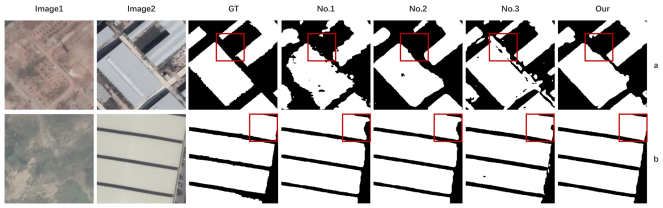

Fig. 14. GZ-CD dataset's ablation study outcomes.

dataset, with two sets of samples selected to showcase significant edge features. The results revealed that experiment No.1 failed to accurately identify the integrity of target areas in the first sample and performed poorly in edge recognition in the second sample. For Experiment No.2, both samples exhibited evident mistakes in edge judgment, displaying significant recognition errors. In contrast, Experiment No.3 showed better performance in the second sample, accurately capturing edge information, though it also encountered edge detection errors in the first sample.

In the series of ablation experiments conducted on the TIDEM, MSGA, and DE modules, we observed that the removal of any module significantly reduced the model's performance across three datasets, specifically manifested in the decline of F1 scores and IoU metrics. Comparative analyses in Fig. 12–13 illustrate how the removal of these key modules greatly diminished the output quality of the model. Specifically, the absence of the TIDEM module weakened the model's capability to extract and enhance image change information, resulting in substantial information loss; the lack of the MSGA module affected the model's clarity and precision in identifying edge details, showing inadequacies in local detail representation; and the removal of the DE module hindered the model's ability to effectively fuse information across different levels, leading to missed detections. These comprehensive experimental results underscore the importance of each module in enhancing the model's ability to capture image features, process information fusion, and maintain the integrity of key information, confirming their indispensable role in improving model performance.

V. Conclusion

In this article, we introduced an innovative TIMF-Net aimed at enhancing the CD capability in bitemporal remote sensing imagery. We meticulously designed three key modules: the TIDEM module to strengthen temporal interaction information within images, the MSGA module to achieve pixel-level comprehensive integration at the global channel and multiscale levels, and the DE module to perform multilevel feature map fusion on spatial coordinates. Working together, these modules not only efficiently extract and enhance change signals in images but also deeply integrate information across the temporal dimension through precise weighting strategies. TIMF-Net exhibited outstanding performance on three benchmark datasets, achieving industry-leading metrics in F1 scores and IoU, showcasing its superior generalization capability. Nonetheless, we also recognize that the model has not yet reached optimal performance in certain aspects. In the future, we plan to further enhance the model's ability to learn from blurry and subtle feature information and to optimize the network structure. Specifically, we aim to incorporate the concept of diffusion models to better focus the model on targets and improve outcomes while optimizing parameters. This will help advance RSCD technology, achieving more precise surface monitoring and analysis.

References

[1] Z. Lv et al., "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proc. IEEE.*, vol. 110, no. 12, pp. 1976–1991, Dec. 2022.

[2] Z. Lv, F. Wang, G. Cui, and J. Benediktsson, "Spatial-spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412712.

[3] Z. Lv, P. Zhang, W. Sun, J. A. Benediktsson, J. Li, and W. Wang, "Novel adaptive region spectral–spatial features for land cover classification with high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5609412, doi: 10.1109/tgrs.2023.3275753.

[4] Z. Lv, H. Huan, M. Jia, J. Benediktsson, and F. Chen, "Iterative training sample augmentation for enhancing land cover change detection performance with deep learning neural network," *Proc. IEEE.*, vol. 110, no. 12, pp. 1976–1991, Dec. 2022.

[5] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017, doi: 10.1109/tgrs.2017.2707528.

[6] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, Jan. 2016, doi: 10.1109/tgrs.2015.2463075.

[7] A. Saber, I. El-Sayed, M. Rabah, and M. Selim, "Evaluating change detection techniques using remote sensing data: Case study new administrative capital Egypt," *Egyptian J. Remote Sens. Space Sci.*, vol. 24, no. 3, pp. 635–648, Dec. 2021, doi: 10.1016/j.ejrs.2021.03.001.

[8] Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, "Water bodies' mapping from sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the Swir band," *Remote Sens.*, vol. 8, no. 4, Apr. 2016, Art. no. 354, doi: 10.3390/rs8040354.

[9] Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, "Water bodies' mapping from sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the Swir band," *Remote Sens.*, vol. 8, no. 4, 2016, Art. no. 354.

[10] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "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.*, vol. 265, Nov. 2021, Art. no. 112636, doi: 10.1016/j.rse.2021.112636.

[11] Z. Lv, F. Wang, W. Sun, Z. You, N. Falco, and J. A. Benediktsson, "Landslide inventory mapping on VHR images via adaptive region shape similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630211.

[12] M. Hao, W. Shi, H. Zhang, and C. Li, "Unsupervised change detection with expectation-maximization-based level set," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 210–214, Jan. 2014.

[13] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.

[14] M. Davy, F. Desobry, A. Gretton, and C. Doncarli, "An online support vector machine for abnormal events detection," *Signal Process.*, vol. 86, no. 8, pp. 2009–2025, 2006.

[15] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[16] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.

[17] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1688.

[18] W. Zhao, L. Mou, J. Chen, Y. Bo, and W. J. Emery, "Incorporating metric learning and adversarial network for seasonal invariant change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2720–2731, Apr. 2020.

[19] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[20] D. He, Q. Shi, J. Xue, P. M. Atkinson, and X. Liu, "Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113884.

[21] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[22] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[23] M. R. S. B. DATA, "Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing Big Data in earth observation," *Innovation*, vol. 2, no. 1, 2024, Art. no. 100055.

[24] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[25] C. Li et al., "CasFormer: Cascaded transformers for fusion-aware computational hyperspectral imaging," *Inf. Fusion*, vol. 108, 2024, Art. no. 102408.

[26] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[28] W. Wang, L. Liu, T. Zhang, J. Shen, J. Wang, and J. Li, "Hyper-ES2T: Efficient spatial–spectral transformer for the classification of hyperspectral remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, 2022, Art. no. 103005.

[29] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[30] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416.

[31] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628711.

[32] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[34] H. Mahmoudzadeh, "Digital change detection using remotely sensed data for monitoring green space destruction in Tabriz," *Int. J. Environ. Res.*, vol. 1, no. 1, pp. 35–411735, 2007.

[35] P. Coppin, E. Lambin, I. Jonckheere, and B. Muys, "Digital change detection methods in natural ecosystem monitoring: A review," *Anal. Multi-Temporal Remote Sens. Images*, vol. 25, pp. 3–36, 2002.

[36] P. H. Swain and S. M. Davis, "Remote sensing: The quantitative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 3, no. 6, pp. 713–714, Nov. 1981.

[37] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size multitemporal images: Application to tsunami-damage assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1658–1670, Jun. 2007.

[38] C. Gómez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 116, pp. 55–72, 2016.

[39] Y.-W. Liu and H. Chen, "A fast and efficient change-point detection framework based on approximate k-nearest neighbor graphs," *IEEE Trans. Signal Process.*, vol. 70, pp. 1976–1986, 2022.

[40] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[41] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[42] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[43] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[44] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[45] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[46] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[47] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[48] Z. Lin et al., "A structured self-attentive sentence embedding," 2017, *arXiv:1703.03130*.

[49] J. Yuan, L. Wang, and S. Cheng, "STransUNet: A Siamese transunet-based remote sensing image change detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9241–9253, 2022.

[50] X. Xu, J. Li, and Z. Chen, "TCIANet: Transformer-based context information aggregation network for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1951–1971, 2023.

[51] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410213.

[52] 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, 2020, Art. no. 1662.

[53] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[54] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[55] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[56] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.

[57] 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, 2022, Art. no. 8007805.

[58] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1691–1708.

[59] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.

[60] J. Zhang et al., "AERNet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617116.

[61] H. Chen, X. Xu, and F. Pu, "SRC-Net: Bi-temporal spatial relationship concerned network for change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 11339–11351, 2024.

[62] M. Weber et al., "DeepLab2: A tensorflow library for deep labeling," 2021, *arXiv:2106.09748*.

[63] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.