IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 November 2023, accepted 2 January 2024, date of publication 5 January 2024,
date of current version 11 January 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3350173*

RESEARCH ARTICLE

# VisionTwinNet: Gated Clarity Enhancement Paired With Light-Robust CD Transformers

**TIANAO CHEN**[1] **AND AOTIAN CHEN**[2]
[1]Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48105, USA
[2]Electrical and Computer Engineering Department, Georgia Institute of Technology, Atlanta, GA 30332, USA

Corresponding author: Aotian Chen (achen653@gatech.edu)

**ABSTRACT** Deep learning has shown superiority in change detection (CD) tasks, notably the Transformer architecture with its self-attention mechanism, capturing long-range dependencies and outperforming traditional models. This capability provides the Transformer with significant advantages in capturing global-level features of complex changes in objects within high-resolution remote sensing images. Though Transformers are mature in Natural Language Processing (NLP), their application in computer vision, particularly CD tasks, is nascent. Current research on leveraging Transformers for CD reveals limitations, especially under varied lighting and seasonal changes. To address this, we propose VisionTwinNet, a two-stage strategy. First, our Gated EnhanceClearNet, a specially designed deep network reduces image noise and enhances brightness, preserving shadows and correcting color distortions. With its unique gating mechanism, this network can adaptively adjust the importance of features, thereby exhibiting superior performance in various remote sensing image degradation issues. Secondly, we have developed Hybrid Light-Robust CDNet, a hybrid robust lightweight network custom-designed for CD in remote sensing images. This module deeply integrates the advantages of CNN and Transformer and introduces an innovative attention mechanism design, optimizing the key/value dimensions separately, instead of adopting traditional single linear transformations, ensuring efficient detection. Specifically, the LR-Transformer Block employs a lightweight multi-head self-attention mechanism, optimizing computational efficiency while providing richer feature representations. Comparative studies with six CD methods on three public datasets validate VisionTwinNet's robustness and efficacy. Our approach notably reduces algorithmic complexity and enhances the efficiency of the model.

**INDEX TERMS** Automatically adjustable framework, change detection, deep learning, multi-scale feature extraction, transformer.

## I. INTRODUCTION

Change Detection (CD) denotes the comparison and analysis of remote sensing images through multiple time periods. Although definitions vary with application scenarios, CD primarily identifies changes in surface features at the pixel, semantic, or scene levels. This technique has a wide range of applications, including resource management, urban planning, disaster assessment and other areas.

Over time, CD started with pixel-level statistical techniques, like image transformation methods [1], [2], [3], during the 1980s. These early methods were more suited to low and medium-resolution images due to their inherent accuracy constraints. The advent of machine learning and deep learning in the 1990s heralded a leap in CD's accuracy. Initial forays into this domain saw the application of artificial neural networks (ANN) [4] and decision trees [5] to remote sensing images. With these advancements, the focus shifted towards applying machine learning to CD, with notable milestones like the incorporation of Support Vector Machines (SVMs) by Cortes and Vapnik in 1995 [6]. SVMs, offering an optimal balance between model intricacy and learning prowess, surpassed traditional methods in generalization. Additionally, Markov Random Field (MRF) enhanced object-level change detection by merging spatial and spectral information [7], offering an efficient CD technique.

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

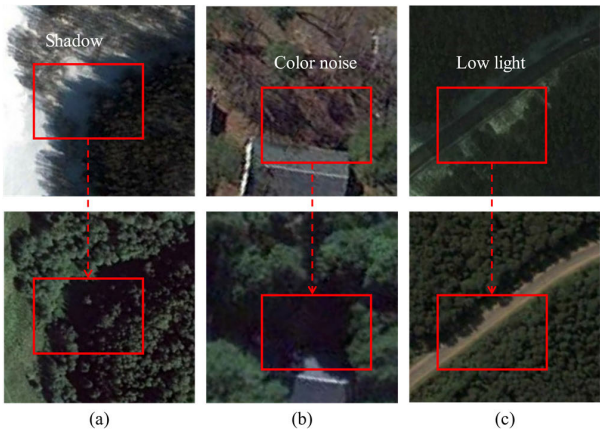


**FIGURE 1. Various factors affecting change detection results. (a) Shadows on tree changes. (b) Color noise error. (c) Low light causes edge blur.**

Challenges in early CD, especially in intricate scenarios, were mitigated with the 2010 emergence of convolutional neural networks (CNNs) and wdeep neural networks [8]. These networks excel in extracting sophisticated scene features and demonstrate resilience to noise. A pivotal moment came in 2012 when Hinton and Alex unveiled the AlexNet for the ImageNet competition [9]. AlexNet's groundbreaking features, like ReLUs and Dropout, have set precedents in CD. The continuous quest for broader information capture led to the integration of attention mechanisms and additional convolutional layers [10].

In recent studies, Transformers have been identified as highly effective for bi-temporal CD models [11]. Compared to CNNs and RNNs, Transformers demonstrate superior discriminative capabilities and an expanded receptive field in the CD domain. Their ability to effectively model within the spatial-temporal domain underscores their proficiency in capturing complex changes within bitemporal images [11]. Bandara's proposal of a Siamese network, which employs a transformer encoder coupled with an MLP decoder, further underscores the pivotal role these architectures play in advancing CD methodologies [12].

At present, various CD algorithms have achieved high precision on mainstream datasets such as DSIFN, LEVIR, and WHU-CD [13], [14], [15], [16]. However, these algorithms exhibit limitations when dealing with bi-temporal images under varying illumination conditions and noise. Within the realm of bi-temporal image CD, identifying color noise and object edge shadows in low-light images remains a significant challenge, as illustrated in Figure. 1. Additionally, current models still lack sufficient detection accuracy regarding the decay of buildings. With the increase in model complexity, computational demands also rise, potentially limiting their utility in real-world applications. Consequently, traditional deep learning methods may exhibit substantial limitations in this application scenario.

To address these issues, this paper proposes a two-stage strategy, VisionTwinNet, that integrates image enhancement and change detection, exhibiting robustness especially for images under low and uneven lighting conditions. Our method not only elevates the visual quality of low-light images but also augments their realism and accuracy in replicating real-world scenes. Specifically, we initially employ Gated EnhanceClearNet for image enhancement, a network comprising two main components: the Illumination Assessment Module and the EnhanceClearNet Module, deciding whether to undertake the image enhancement task through a gating mechanism. With this design, our network can adaptively handle input images of varying quality, enhancing and denoising them only when necessary. This approach not only improves the model's flexibility and efficiency but also maintains the images' original characteristics, thereby improving overall performance without compromising visual quality. Subsequently, a Hybrid Light-Robust CDNet, a mix of robust lightweight network, is utilized for change detection tasks. The overall architecture of this network mainly consists of Local Feature Capture Unit (LFCU), LR-Transformer Block, and Scale-Adaptive Decoder, which collaborate to achieve efficient detection of changes in remote sensing images. From extracting local features to integrating multi-scale features and then generating high-resolution prediction results, the method ensures computational efficiency and effective local feature extraction. Moreover, the LR-Transformer Block designed in this paper not only processes multiple different attention heads in parallel in different subspaces, enhancing the model's representational capability, but its innovative design for the attention mechanism also optimizes computational efficiency and alleviates computational burden, enabling higher efficiency and performance when handling large-scale input features.

The main contributions of our work can be summarized as follows.

1) Inspired by the Retinex theory [17], we propose Gated EnhanceClearNet, a lightweight image enhancement network. This network first assesses the illumination quality of the input image and then determines whether to perform image enhancement. This adaptive enhancement strategy not only improves the model's efficiency and flexibility but also ensures the improvement of overall performance without compromising visual quality.
2) After the image enhancement stage, we introduce the Hybrid Light-Robust CDNet for CD tasks. This module is specifically designed for images under low and uneven illumination conditions, thereby enabling it to effectively detect changes within images even under intricate lighting circumstances. This module deeply integrates the advantages of CNN and Transformer and introduces an innovative attention mechanism design, optimizing the key/value (k/v) dimensions separately, instead of adopting traditional single linear transformations, effectively reducing computational complexity.
3) By integrating Gated EnhanceClearNet and Hybrid Light-Robust CDNet, this paper proposes a two-stage

strategy named VisionTwinNet. Initially, Gated EnhanceClearNet determines whether image enhancement is necessary, followed by Hybrid Light-Robust CDNet conducting the CD task. This structure ensures that our method provides high quality results in a wide range of lighting conditions.

4) We conducted experiments on four publicly available datasets, and the results show that our method outperforms existing methods under both low-light and uneven illumination conditions. This further proves the effectiveness and robustness of our method. Meanwhile, the method proposed in this paper effectively reduces the algorithm complexity and ensures higher efficiency of the model.

The rest of this paper is organized as follows. Section II introduces the related work. Section III describes the proposed Gated EnhanceClearNet and Hybrid Light-Robust CDNet in detail. Section IV presents the experimental results and conducts a comparative analysis. Section V gives the conclusion.

## II. RELATED WORK

The field of change detection (CD) gained significant attention was due to the introduction of traditional machine learning methods, such as supporting vector machines (SVMs) (Vapink,1998) decision trees (DTs) [18], and random forests (RFs) [5]. SVMs, for example, can effectively detect changes in images by mitigating the uncertainty associated with change thresholds, exhibiting strong generalization capabilities, and can efficiently processing data in high-dimensional feature spaces [1]. In contrast to traditional methods such as direct comparison and image transformation methods, which are limited to detecting changes in medium to low-resolution satellite images, DTs can effectively detect changes in high spatial resolution multispectral data and eliminate noise related to shadows in the image [4]. RFs have demonstrated similar capabilities and have subsequently played a crucial role in the reimplementation of direct comparison methods in high-resolution remote sensing imagery [19], [20]. Markov Random Fields (MRFs) have also been widely used in CD research [21]. Despite continuous optimization [7], some limitations remain, such as insufficient preservation of change edge details, resulting in excessive smoothing [19]. This poses a significant challenge for CD in high-resolution remote sensing imagery. To address this issue, Im and Jensen proposed an unsupervised CD method that combines MRFs with an edge information-based penalty function to avoid excessive smoothing and improve the generation of difference maps [7]. During the same period, other improved direct comparison methods, such as conditional random fields (CRFs), were also applied to CD in high-resolution remote sensing imagery [22], [23].

During this period, traditional machine learning and other methods encountered many limitations in the field of CD [24]. The emergence of deep learning provided unprecedented opportunities for advancement in this area, with convolutional neural networks (CNNs) for CD tasks in various fields. For example, an innovative CNN-based model for detecting changes in synthetic aperture radar (SAR) imagery was introduced in recent research [25]. Another study proposed three fully convolutional neural network architectures and two Siamese extensions of fully convolutional networks, collectively known as FC-EF, for performing CD using a pair of images [26]. Due to the limitations of existing algorithms in terms of feature representation and discriminative ability, attention mechanisms have been introduced. Liu proposed a deep Siamese convolutional network with dual-task constraints (DTCDSCN) to address the CD problem, effectively addressing these issues [26]. To enhance the completeness of object boundaries and the compactness of their interiors in the generated change maps, methods that use attention mechanisms for change map reconstruction have been introduced [27].

Deep learning networks have demonstrated great potential in CD tasks [28], [29], [30]. However, the Transformer architecture is increasingly being widely applied to CD tasks because of its capability to model dependencies with long-range and learn more discriminative global-level features. This is particularly effective for handling the complex changes of objects in high-resolution remote sensing CD. Chen proposed a bi-temporal image transformer (BIT) that expresses bi-temporal high-resolution images into semantic tokens and uses a transformer encoder to efficiently model spatial-temporal contexts for CD [31]. Compared to CNNs, this method has lower computational costs and fewer model parameters [32]. In 2022, a Siamese network architecture that features a transformer encoder with hierarchical structure and a lightweight MLP decoder was introduced [33], which achieved better performance in edge segmentation and shadow processing than previous counterparts.

However, the recognition of color noise and object edge shadows in bi-temporal image CD, particularly in low-light images, remains a significant challenge. Chen suggests that variations in lighting conditions between images taken at different periods can result in the identical object appearing with different colors and brightness levels. This phenomenon can lead to erroneous CD results. Hence, it is essential to perform color correction and uniform balance of lighting conditions [34]. Furthermore, compared to CNNs, the transformer employs attention as an alternative to convolution to realize global context modeling [31]. However, since the original attention mechanism calculates pairwise feature affinities at all spatial locations, it entails high computational burden and substantial memory occupation, especially for high-resolution inputs. Consequently, there is a necessity to explore more efficient attention mechanisms. Unlike previous work, we design a two-stage approach that combines image enhancement and change detection, specifically addressing images under low and uneven lighting conditions. The method proposed exhibits significant robustness. Our method not only improves the visual quality of low-light images, but

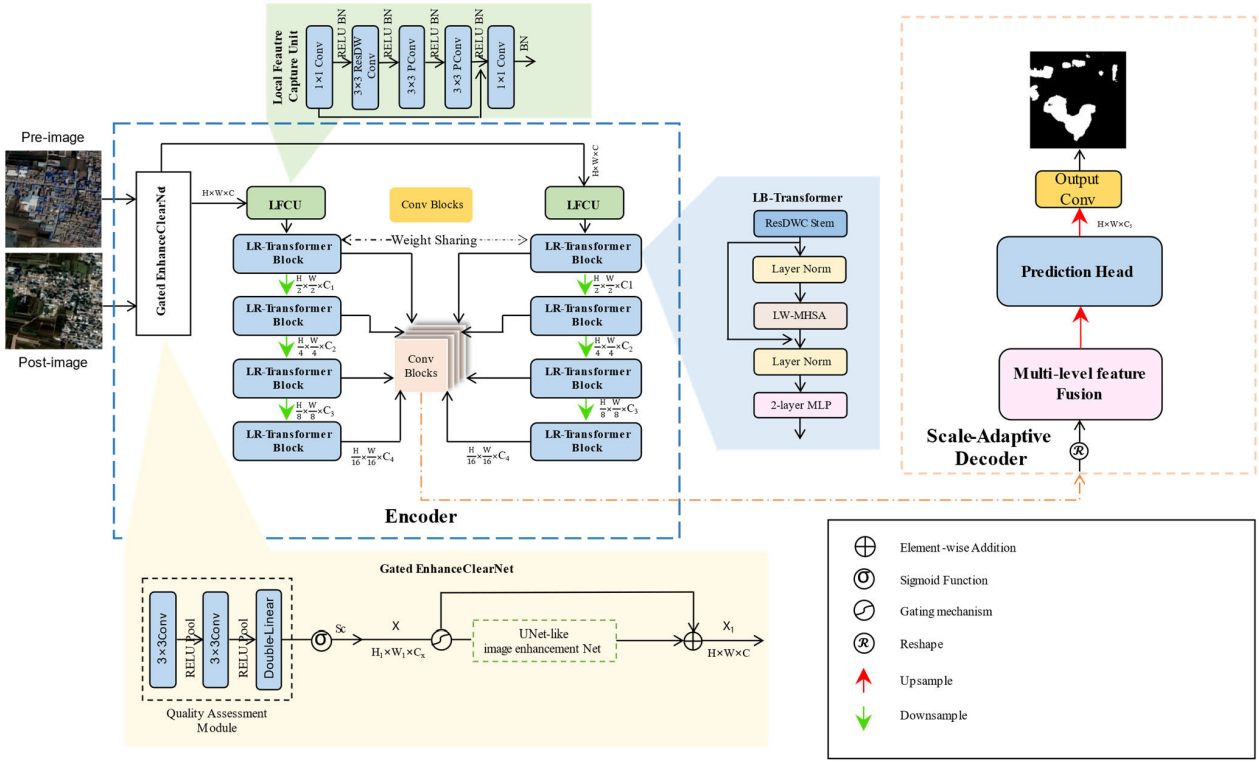


**FIGURE 2.** Overall structure of our proposed model VisionTwinNet.

also augments their realism and accuracy in relation to real-world scenarios. The introduction of the gating mechanism enables our network to adaptively process input images with different qualities, which not only improves the flexibility and efficiency of the model, but also helps to maintain the original characteristics of the images. In addition, the innovative design of the attention mechanism in this paper effectively reduces the computational complexity of the model.

## III. METHOD

In this section, we provide a detailed explanation of the proposed VisionTwinNet architecture, designed for change detection (CD) tasks, and its application to remote sensing images. The main objective of this work is to address the issues of model performance and efficiency in situations where images are under complex lighting conditions. The overall flow of our VisionTwinNet architecture is illustrated in Figure. 2. This VisionTwinNet architecture employs a two-stage strategy, mainly including two modules, Gated EnhanceClearNet and Hybrid Light-Robust CDNet. Gated EnhanceClearNet determines whether image enhancement is necessary, and subsequently, Hybrid Light-Robust CDNet conducts the CD task. This structure ensures that our method provides high-quality results in a wide range of lighting conditions. More specifically, given an input bi-temporal image, the Gated EnhanceClearNet generates the enhanced images. It includes the use of gating mechanisms and quality assessment modules to flexibly determine whether image enhancement and degradation removal are needed. The Decomposer divides the image into reflectance and illumination. The EnhanceNet improves image lighting, and the ClearNet removes image degradation and artifact issues. Subsequently, the enhanced images are fed into the Local Feature Capture Unit (LFCU) to extract local features. These features then go through four consecutive LR-Transformer Blocks to compute feature differences. Finally, a Scale-Adaptive Decoder effectively fuses features of different scales and generates high-resolution prediction results, as illustrated in Figure. 3.

### A. GATED ENHANCECLEARNET

Inspired by the Retinex theory, this paper proposes a simple and effective light-weight network, Gated EnhanceClearNet, as illustrated in Figure. 4. This network is composed of two main components: the Illumination Assessment Module and the EnhanceClearNet Module, with a gating mechanism deciding whether to perform image enhancement tasks.

Initially, the input image enters the Illumination Assessment Module, where a convolutional neural network (Quality Assessment Module) evaluates the illumination quality of the image. The Quality Assessment Module comprises three convolution layers, a maximum pooling layer, and two fully connected layers. The last layer uses a Sigmoid activation function to map the output to a value between 0 and 1,

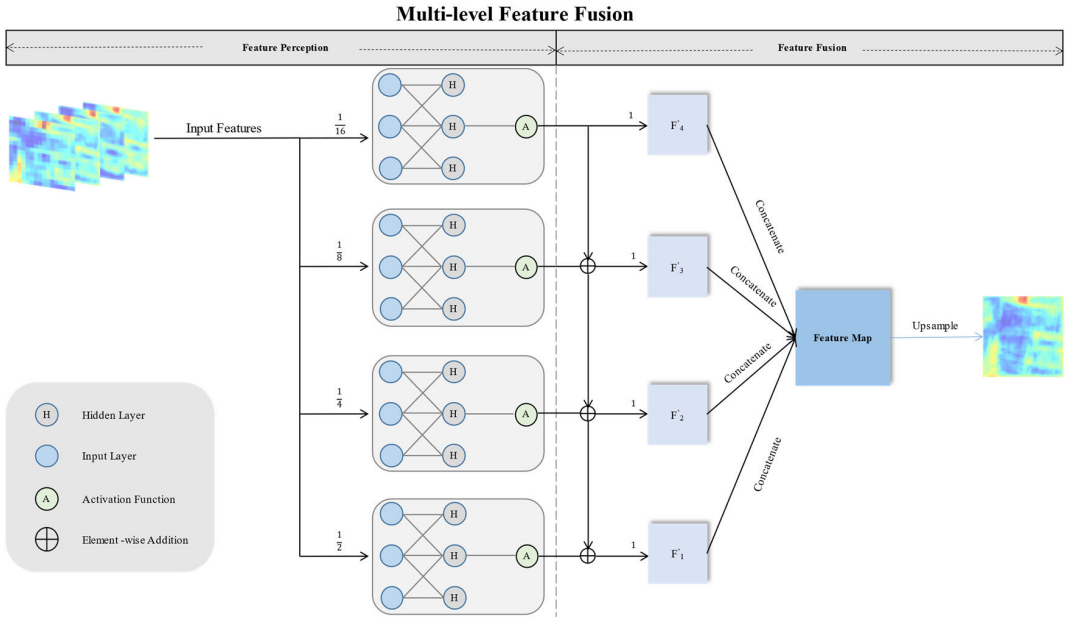


**FIGURE 3.** Multi-level feature fusion module.

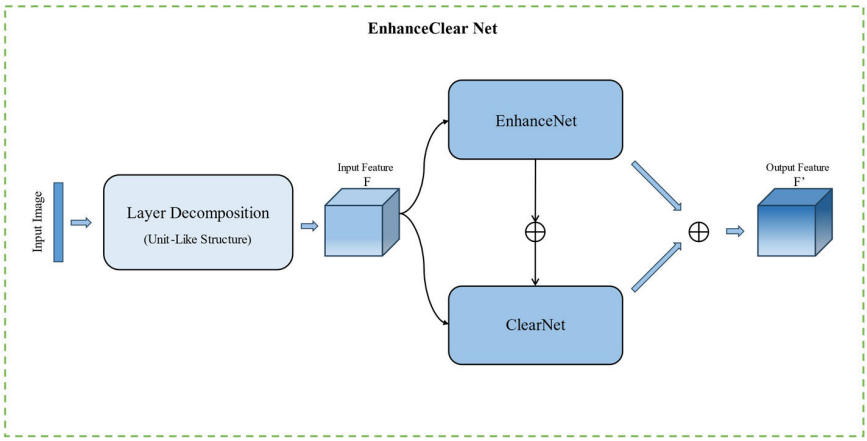


**FIGURE 4.** EnhanceClearNet.

representing the quality score of the image. Subsequently, based on the quality score from the Quality Assessment Module, the gating mechanism decides whether the image requires enhancement. If the gate value is below a certain threshold, the image is passed to the EnhanceClearNet Module for further enhancement.

EnhanceClearNet is a nested structure consisting of two sub-networks. The primary network receives the original image as input and inspired by the Retinex theory, decomposes it into reflectance and illumination component. This decomposition allows the main network to handle the reflection and illumination information in the image separately. Firstly, the main network processes the illumination component through a sub-network called ''EnhanceNet'' to improve or enhance the brightness and overall lighting conditions of the image. ''EnhanceNet'' is responsible for adjusting the image's brightness, contrast, etc., effectively enhancing the visual quality of the image without introducing unnecessary alterations, making the image brighter, clearer, and highlighting the detail information in the image. Secondly, the main network processes the reflectance component through another sub-network called ''ClearNet'' to remove or reduce degradations and artifacts in the image caused by various factors such as rain, smog, or other visual distortions. ''ClearNet'' focuses on denoising and removing unnecessary components from the image, thereby enhancing the clarity and visibility of the image. In the output phase of the main network, the outputs of these two sub-networks are recombined to produce the enhanced image. Through such a nested structure, EnhanceClearNet is capable of both

enhancing illumination and removing image degradation, producing clearer images with better visual effects. The entire network architecture is designed to be end-to-end, learning, and extracting reflectance and illumination information directly from the original images to better accomplish the image enhancement tasks.

#### 1) QUALITY ASSESSMENT & GATING MECHANISM

To enhance the adaptability and performance of the model to images of varying quality, this paper introduces an adaptive image enhancement strategy based on a gating mechanism, as illustrated in Figure. 2. The core idea of this strategy is to automatically determine whether image enhancement and degradation removal are necessary, based on the quality of the input image.

Firstly, we have designed a Quality Assessment module, consisting of a series of convolution layers, to estimate the quality of the input image. This module outputs a scalar between 0 and 1, representing the quality score of the image. Subsequently, we have introduced a gating mechanism that controls whether the image passes through the enhancement module (EnhanceClearNet), based on the output from the Quality Assessment module. Specifically, the gating mechanism transforms the quality score into a binary decision signal through a threshold function. If the image quality score is above a predefined threshold (close to 1), the image bypasses the enhancement module and directly enters the next stage of the network; if the quality score is below the threshold (close to 0), the image will be processed by the enhancement module. Finally, the enhanced image (if enhanced) or the original image (if not enhanced) is passed to the next layer of the network for further processing.

With this design, our network can adaptively handle input images of different qualities, enhancing images and removing degradations only when necessary. This not only improves the flexibility and efficiency of the model but also helps to maintain the original characteristics of the images, thus enhancing overall performance without sacrificing visual quality.

#### 2) DECOMPOSER

The Decomposer module is a core part of EnhanceClearNet, as illustrated in Figure. 4, responsible for effectively decomposing the input image into its reflectance and illumination. This module uses the Leaky ReLU function with a slope of 0.2 as the activation function and adopts bilinear interpolation for upsampling, combined with corresponding feature maps to ensure the consistency of spatial dimensions. By introducing three residual blocks, this module enhances the information flow, helping to reduce the vanishing gradient problem during training. These designs allow each sub-network to focus on their respective tasks, improving the flexibility and efficiency of the model. The mathematical formula for the Leaky ReLU function can be presented by

$$f(x) = \max(ax, x) \tag{1}$$

*Reflectance branch:* Its main task is to accurately capture the reflective properties of the image. The branch follows a UNet-like structure [35], with the inclusion of additional components such as an improved ReLU activation function, residual blocks, extra convolution layers. The proposed branch employs a down-sampling technique using max pooling, similar to the UNet architecture, and an up-sampling technique using interpolation functions, which adopts bilinear interpolation. A key feature of this branch is its mathematical formula, which, by defining how to calculate weighted average pixel values and perform horizontal linear interpolation, can flexibly handle tensors of different sizes and maintain the consistency of spatial dimensions [36]. The mathematical formulate can be presented by

$$R_1 = I(xx, y) \tag{2}$$

$$R_1 = \frac{I(x_1, y_1) \times (x_2 - \mathrm{x}')}{(x_2 - x_1)} + \frac{I(x_2, y_1) \times (\mathrm{x}' - x_1)}{(x_2 - x_1)} \tag{3}$$

$$R_2 = \frac{I(x_1, y_2) \times (x_2 - \mathrm{x}')}{(x_2 - x_1)} + \frac{I(x_2, y_2) \times (\mathrm{x}' - x_1)}{(x_2 - x_1)} \tag{4}$$

$$I(x', y') = \frac{R_1 \times (y_2 - y')}{(y_2 - y_1)} + \frac{R_2 \times (y' - y_1)}{(y_2 - y_1)} \tag{5}$$

where $I(x', y')$ is a weighted average of the surrounding pixel values using the bilinear interpolation formula, and $R_1, R_2$ represent the results of horizontal linear interpolation on the image values at $y_1$ and $y_2$, respectively.

*Illumination branch:* Illumination branch. It is another crucial component of EnhanceClearNet, focusing on precise modeling of the illumination component. This branch employs two convolution layers and ReLU activation functions as illumination layers and generates illumination output using a $1 \times 1$ convolution layer and a Sigmoid activation function. With this design, the Illumination branch is capable not only of improving overall visual quality and lighting conditions but also of ensuring accuracy and robustness under various lighting scenarios.

#### 3) ENHANCENET

Compared to bright-light images, low-light images are more prone to degradation and disturbances. EnhanceNet, as illustrated in Figure. 4, is a solution targeted at issues of degradation and disturbance in low-light images, aiming to improve output quality by enhancing the reflectance map [37]. Several key sections are specifically designed in this network structure to achieve this goal.

Firstly, EnhanceNet employs a novel reflectance recovery network, which incorporates both illumination information and degraded reflectance to produce higher-quality outputs under challenging low-light conditions. Building on this, EnhanceNet introduces a Multi-Scale Attention Module (MSIA), a design inspired by KinD++ [38]. The MSIA module performs well in resolving color distortions and overexposure in low-light images. However, relying on MSIA

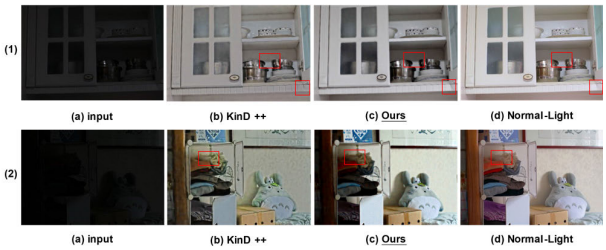


**FIGURE 5.** Visual comparison with KinD++ in LOL dataset.

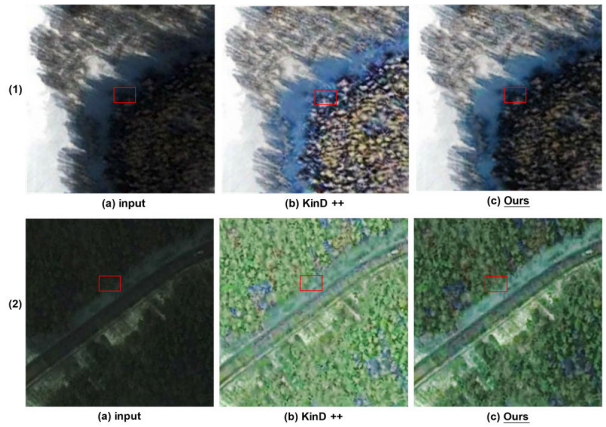


**FIGURE 6.** Visual comparison with KinD++ in LEVIR-CD dataset.

alone is not sufficient, and hence, EnhanceNet adds squeeze-and-excitation (SE) blocks after each MSIA module to adaptively recalibrate channel-wise feature responses in the convolutional neural network, enhancing its representational ability.

Additionally, to ensure the effective learning and updating of weights during the training process, EnhanceNet incorporates three residual blocks into its architecture. These combined designs enable the model to proficiently preserve texture details in dark areas (such as frosted glass) and yield noticeably better results in this aspect compared to the original algorithm, as shown in Figure. 5(1).

Beyond the features mentioned above, EnhanceNet can also more effectively recover color information in dim-light images, rendering the colors of the restored images closer to those of the original scenes, as demonstrated in Figure. 5(2). Meanwhile, the model efficiently suppresses overexposure in bright areas of low-light images on high-precision remote sensing datasets like LEVIR, as depicted in Figure. 6.

In summary, by adopting and enhancing the MSIA module from KinD++ and introducing other innovative techniques, EnhanceNet provides an effective solution for the restoration and enhancement of low-light images. This design not only elevates the visual quality of low-light images but also augments their realism and accuracy in reflecting real-world scenarios.

### 4) CLEARNET

ClearNet is a specialized neural network architecture, as illustrated in Figure. 4, developed for the purpose of manipulating the illumination intensity within an input image. This process aims to balance the lighting contrast across different areas of the image, creating a more visually pleasing and natural result. By connecting the input image and input ratios, three consecutive convolutional layers in this network architecture—each utilizing $3 \times 3$ filters, a stride of 1, and Leaky ReLU activation functions—are responsible for extracting features related to illumination intensity within the image. The subsequent convolutional layer, employing $3 \times 3$ filters and a Sigmoid activation function, transforms these features into representations associated with the desired illumination modifications.

This approach allows the enhancement of the brightness of darker areas by adding a smaller proportion of light while maintaining the consistency of the original image with real-world lighting conditions and increasing the brightness of bright areas with an equal or greater amount of light. To better adapt to multiple datasets, the ClearNet structure has also undergone meticulous parameter tuning. This design can eliminate or reduce degradations and artifacts in the image caused by various factors, thus enhancing the clarity and visibility of the image.

### 5) LOSS FUNCTION

The loss function of the Gated EnhanceClearNet is composed of three parts, corresponding to the Decomposer network, the EnhanceNet network, and the ClearNet network, respectively. Loss function for Decomposer network is presented by

$$L_{Dec} = L^H_{rec} + L^L_{rec} + 0.009L^R_{equal} + 0.2L^H_{mutual} + 0.15L^H_{input} + 0.15L^L_{input} \tag{6}$$

where $L^H_{rec}$ and $L^L_{rec}$ compute the reconstruction loss under high and low illumination conditions, respectively; $L^R_{equal}$ computes the mean absolute difference between predicted reflectance under low and high illumination conditions; $L^H_{mutual}$ calculates the mutual illumination conditions; $L^H_{input}$ and $L^L_{input}$ compute the mutual illumination loss between the predicted illumination and the input image under high and low illumination conditions, respectively.

Loss function for EnhanceNet can be given by

$$L_{reflec} = L_{square} + L_{ssim\&se} \tag{7}$$

where $L_{square}$ calculates the mean squared error (MSE) between the predicted reflectance and the actual reflectance, and $L_{ssim\&se}$ calculates the SSIM [39] loss between the predicted and actual reflectance.

Loss function for ClearNet can be given by

$$L_{illu} = L_{square} + L_{grad} \tag{8}$$

where $L_{grad}$ calculates the gradient loss between the predicted and ground truth illumination maps, and $L_{square}$ computes the MSE between the predicted and actual illumination values.

### B. HYBRID LIGHT-ROBUST CDNET

The Hybrid Light-Robust CDNet proposed in this paper is a hybrid robust lightweight network specifically designed for change detection in remote sensing images. The overall architecture of this network mainly consists of three primary components: Local Feature Capture Unit (LFCU), LR-Transformer Block, and Scale-Adaptive Decoder, which work together to efficiently detect changes in remote sensing images.

Firstly, the input image is fed into the Local Feature Capture Unit (LFCU), a specially designed module for efficiently and robustly extracting spatial and local features. LFCU precisely controls the localization of convolution operations through partial convolution and $1 \times 1$ convolution blocks, as well as residual connections, thereby effectively extracting local features while maintaining computational efficiency.

Subsequently, the output from the LFCU is fed into the LR-Transformer Block. This block initially utilizes the Residual Depthwise Convolution (ResDWC) module for depthwise separable convolution and residual connections, enhancing computational efficiency, model expressiveness, and training stability. Then, it implements a Lightweight Multi-Head Self-Attention mechanism, allowing parallel processing of multiple different attention heads in different subspaces, enhancing the model's representational capability. The innovative attention mechanism design separately optimizes the key/value (k/v) dimensions, significantly improving computational efficiency and reducing computational burden. Moreover, the LR-Transformer Block also introduces a 2-layer MLP module to transform the features at each location, allowing the model to learn and capture more complex feature representations and improve its performance.

Finally, the output from the LR-Transformer Block is decoded by the Scale-Adaptive Decoder. This decoder employs a strategy of multi-scale feature fusion and adaptive feature selection to address the issues of low detection accuracy and high false alarm rates in traditional decoders due to scale variations and complex backgrounds when dealing with change detection tasks in remote sensing images. The Scale-Adaptive Decoder, through the MLP decoder head, convolution difference module, intermediate prediction module, final linear fusion layer, upsampling convolution layer, residual block, and final prediction head, accomplishes effective fusion of features at different scales and generates high-resolution prediction results.

In conclusion, through the collaborative effort of these three main components, the Hybrid Light-Robust CDNet achieves efficient detection of changes in remote sensing images. From local feature extraction to multi-scale feature fusion, and finally to the generation of high-resolution prediction results, each part plays a crucial role within the entire network, collectively forming a powerful network for change detection in remote sensing images.

#### 1) LOCAL FEATURE CAPTURE UNIT

This paper proposes a module named Local Feature Capture Unit (LFCU) designed to extract spatial and local features efficiently and robustly through meticulously crafted structures and operations. The core components of LFCU are partial convolution and $1 \times 1$ convolution blocks, working together to capture local features and spatial context.

Partial convolution proposes an innovative convolution strategy, its main advantage being significantly improving computational efficiency and effectively extracting local features. By applying convolution only over a subset of the input channels, partial convolution not only reduces computation and memory access, enabling the model to handle larger inputs or deeper network structures, but also captures significant spatial features of the input, retaining the original information of the other channels. Moreover, partial convolution enhances the model's expressiveness by altering the way convolution is applied, enabling the model to learn more complex feature representations without increasing computational burden. Given an input feature map X, we can divide it into two parts as follows:

$$X_1, X_2 = \text{split}(X, dim_{conv}) \tag{9}$$

where $dim_{conv}$ is the number of channels to apply convolution, $X_1$ are the selected channels, $X_2$ are the channels that are kept unchanged. Then, convolution operations are applied to $X_1$, represented as:

$$Y_1 = \text{Conv}(X_1, W) \tag{10}$$

where $W$ represents the convolution kernel weights. The final output feature map is given by the following concatenation operation:

$$Y = \text{Concat}(Y_1, X_2) \tag{11}$$

The introduction of this partial convolution reduces computation and memory access, thereby capturing important local features while maintaining computational efficiency. Overall, the operation of partial convolution can be represented as the following composite function:

$$\begin{aligned} Y = \text{Concat}(&\text{Conv}(\text{split}(X, dim_{conv})[0], W), \\ &\text{split}(X, dim_{conv})[1]) \end{aligned} \tag{12}$$

Moreover, $1 \times 1$ convolution blocks allow the channel number to be altered without changing spatial dimensions, thereby achieving channel integration and expansion. This design aids in enhancing the expressive capability of the model while maintaining computational efficiency. LFCU also includes a residual connection, enhancing training stability and convergence speed, helping to retain the original information of the input, thus enhancing the robustness of the model.

In summary, the LFCU module, by precisely controlling the localization of convolution operations and combining partial convolution, $1 \times 1$ convolution blocks, and residual connections, addresses the challenge of effectively extracting local features while maintaining computational efficiency.

This integrated design not only captures spatial context but also emphasizes the importance of local features, representing an efficient and robust method of feature extraction.

### 2) LR-TRANSFORMER BLOCK

The LR-Transformer Block is designed to address a series of challenges in remote sensing image change detection, integrating innovative designs and multi-scale strategies. This block incorporates the Residual Depthwise Convolution (ResDWC) module, a lightweight multi-head self-attention mechanism, and a two-layer Multi-Layer Perceptron (MLP), providing a powerful framework to overcome detection issues in long-distance dependencies and complex backgrounds. The ResDWC module, utilizing depthwise separable convolution and residual connections, enhances the model's expressive ability and computational efficiency while maintaining training stability. The lightweight multi-head self-attention mechanism exhibits exceptional performance when handling large-scale input features, possessing superior computational efficiency and parallel processing capabilities. The two-layer MLP module further enriches the model's feature expressive ability. Moreover, by combining multi-scale feature fusion and adaptive feature selection strategies, the Scale-Adaptive Decoder successfully resolves the problems of low detection accuracy and high false detection rates caused by scale variations and complex backgrounds.

*Residual Depthwise Convolution (ResDWC):* The ResDWC module combines the advantages of depthwise separable convolution and residual connections to enhance computational efficiency, augment the expressive capability of the model, and boost training stability. Depthwise separable convolution [40] significantly reduces computation and model parameters by decomposing the convolution operation into two steps: depth convolution and point convolution. This method allows each input channel to have a corresponding convolution kernel, achieving depth convolution. Residual connections, on the other hand, create a shortcut connection by adding input features to the convolutional features, allowing the model to learn an identity mapping between the input and output, which is very beneficial for training deep networks. Additionally, residual connections can also alleviate the vanishing gradient problem, enabling deeper networks to be effectively trained. Thus, as the first part of the encoder, the ResDWC module provides a powerful and stable foundation for subsequent feature extraction and information transmission.

*Lightweight Multi-Head Self-Attention:* It is designed to efficiently reduce computational complexity while maintaining the performance of the model. This attention mechanism incorporates an innovative design, optimizing the dimensions of keys and values (k/v) separately, instead of using traditional singular linear transformation. This design enhances computational efficiency and reduces computational burden, allowing the self-attention mechanism to exhibit higher efficiency when dealing with large-scale input features. Within the multi-head self-attention mechanism, the input feature space is divided into multiple heads, each performing attention operations independently. Specifically, for the h^th head, the Query (Q), Key (K), and Value (V) are generated through the following linear transformations:

$$Q_h = X_h W_{qh}, \quad K_h = X_h W_{kh}, \quad V_h = X_h W_{vh} \tag{13}$$

where $X_h$ represents the input features of the $h^{th}$ head, $W_{qh}$, $W_{kh}$, $W_{vh}$ represents the corresponding weight matrices, respectively.

Additionally, we incorporate the concept of Relative Position Encoding [41]. This not only enables the capturing of interactions among input features during the computation of attention scores but also enhances the model's capability to apprehend more complex dependencies by providing spatial positional information through the integration of relative position encoding. This, in turn, leads to an elevation in the model's performance. The attention scores are computed as follows:

$$\text{Atten}(Q_h, K_h, V_h) = \text{soft}\max(\frac{Q_h K_h^T}{\sqrt{d_{kh}}} + Relative\_Pos)V_h \tag{14}$$

where $d_{kh}$ represents the dimension of the key in the $h^{th}$ head, and $Relative\_Pos$ represents the relative position encoding.

This lightweight multi-head self-attention mechanism not only achieves parallel processing of multiple distinct attention heads in different subspaces, enhancing the model's representational capability, but also its innovative design optimizes computational efficiency and reduces computational burden. Thus, it exhibits superior efficiency and performance when dealing with large-scale input features.

*2-Layer Multilayer Perceptron (MLP):* Following the self-attention mechanism, this paper introduces a 2-layer MLP module, used for transforming the features at each position. Consequently, the model can learn and capture more complex feature representations, enhancing the model's performance. Specifically, the 2-layer MLP module first processes the input features through a linear transformation and a non-linear activation function (ReLU), and then obtains the output features through a second linear transformation. This process can be represented as:

$$H_1 = \text{ReLU}(XW_1 + b_1), \quad H_2 = H_1 W_2 + b_2 \tag{15}$$

where $X$ represents the input features, $W_1$ and $W_2$ are the weight matrices, $b_1$ and $b_2$ are the bias terms, and $H_1$ and $H_2$ are the output features of the first and second layers, respectively. In this way, the 2-layer MLP module can effectively enhance the expression ability of the model and improve the performance of the model.

### 3) SCALE-ADAPTIVE DECODER

The Scale-Adaptive Decoder employs strategies of multi-scale feature fusion and adaptive feature selection to address the challenges faced by traditional decoders in remote sensing

image change detection tasks, where scale variations and complex backgrounds lead to lower detection accuracy and a higher rate of false detections.

The main components of the Scale-Adaptive Decoder include: an MLP decoder head, a convolutional difference module, an intermediate prediction module, a final linear fusion layer, an upsampling convolutional layer, residual blocks, and a final prediction head. Specifically, the MLP decoder head maps feature at different scales into the unified embedding space. The convolutional difference module computes the feature differences between two time points. The intermediate prediction module generates change predictions at each scale. The final linear fusion layer merges differential features from all scales. The upsampling convolutional layer and the residual blocks work together to produce high-resolution predictions, and the final prediction head generates the ultimate change detection results.

In detail, features from different scales are mapped to a unified embedding space through the MLP decoder head. Then, the feature differences between two time points are computed by the convolutional difference module, capturing the change information in the images. Based on this, change predictions are generated at each scale to understand the image changes from various scales and perspectives. Subsequently, differential features from all scales are merged by the final linear fusion layer, forming a comprehensive feature representation. This representation, encompassing change information from all scales, provides rich contextual information for subsequent predictions. To produce high-resolution prediction results, an upsampling convolutional layer and a residual block are designed. These modules collaboratively upscale the low-resolution feature representations to ensure the spatial details in the prediction results. Finally, the high-resolution feature representations are transformed into the ultimate change detection results by the final prediction head. This step maps the feature representations to the prediction space, generating the desired change detection results. The formula for the Scale-Adaptive Decoder is as follows:

$$\hat{c}_{t1} = \text{MLP}(c_{t1}),\ \hat{c}_{t2} = \text{MLP}(c_{t2}) \tag{16}$$

$$\Delta c = \text{difference}(\hat{c}_{t1}, \hat{c}_{t2}) \tag{17}$$

$$p = \text{prediction}(\Delta c) \tag{18}$$

$$\hat{c} = \text{linear\_fuse}(\Delta c_1, \Delta c_2, \Delta c_3, \Delta c_4) \tag{19}$$

$$y = u(\hat{c}) + r(\hat{c}) \tag{20}$$

$$result = \text{change\_probability}(y) \tag{21}$$

where $\hat{c}_{t1}$ and $\hat{c}_{t2}$ are the features at two time points.

The advantage of the Scale-Adaptive Decoder is that it can adaptively select and fuse features of different scales, thereby dealing more effectively with the scale variations and complex background issues in the task of remote sensing image change detection. Additionally, it employs a strategy of multi-scale prediction and final prediction, allowing the decoder to generate meaningful change predictions at each scale, thereby improving the accuracy and robustness of change detection.

**TABLE 1. Five dataset description.**

| Dataset | Train | Test | Val | Size | Type |
|---|---|---|---|---|---|
| LEVIR-CD | 7120 | 2048 | 1024 | 256*256 | CD |
| DSIFN-CD | 14400 | 192 | 1360 | 256*256 | CD |
| WHU-CD | 3600 | 1200 | 600 | 256*256 | CD |
| L-LEVIR-CD | 7120 | 2048 | 1024 | 256*256 | Manual |
| LOL | 485 | 15 | - | 600*400 | Low-Light |

### C. LOSS FUNCTION

For the CD task, we address the issue of imbalanced data by employing a hybrid loss function. Given the pixel-level nature of our CD task, there's a notable imbalance between the proportions of changed and unchanged pixels. To mitigate this, we've adopted a composite loss function that combines both the weighted cross-entropy (WCE) [42] and the dice coefficient loss (DICE) [43]. This is mathematically expressed as:

$$L_{total} = L_w + L_d \tag{22}$$

$$L_w = -\sum_{i=1}^{H\times W} [w_1 y_i \log(p_i) + w_2(1-y_i)\log(1-p_i)] \tag{23}$$

$$L_{DICE} = 1 - \frac{2\sum_{i=1}^{H\times W} y_i p_i}{\sum_{i=1}^{H\times W} y_i + \sum_{i=1}^{H\times W} p_i} \tag{24}$$

where $L_w$ denotes the weighted cross-entropy loss, $L_d$ denotes the dice coefficient loss, $y_i$ denotes the true label for the pixel, indicating whether it has changed (1) or remained unchanged (0), $w_1$ and $w_2$ denotes the weights of changed and unchanged classes, respectively, $p_i$ denotes the predicted probability of changed pixel.

## IV. EXPERIMENTS AND ANALYSIS

To validate the efficacy of our proposed VisionTwinNet, we conducted experiments on three publicly available change detection (CD) datasets: LEVIR-CD [44], DSIFN-CD [28], and WHU-CD [45], as illustrated in TABLE 1. Specifically, the performance of Gated EnhanceClearNet was independently evaluated on the LOL dataset [46]. In this section, we first describe the dataset used and then compare the CD methods. Subsequently, we elucidate the experimental setup and describe the evaluation metrics for the network. Finally, experimental results and ablation studies are presented.

### A. DATASET DESCRIPTION

LEVIR-CD [44]: This dataset is a newly created, extensive CD dataset in buildings, which consists of 637 high-resolution Google Earth image patch pairs, each with a size of 1024 × 1024 pixels. We divide images of size 1024 × 1024 into non-overlapping patches of size 256 × 256. These patches are distributed into three parts to

**TABLE 2. Quantitative comparison results of different CD methods on the three CD Test Sets.***

| Method | WHU-CD<br>Pre. / Rec. / F1 / IoU / OA | LEVIR-CD<br>Pre. / Rec. / F1 / IoU / OA | DSIFN-CD<br>Pre. / Rec. / F1 / IoU / OA |
|---|---|---|---|
| FC-EF [26] | 71.63 / 67.37 / 69.37 / 53.11 / 97.61 | 84.91 / 74.36 / 79.36 / 64.83 / 97.98 | 72.61 / 52.73 / 61.09 / 43.98 / 88.59 |
| FC-Siam-Di [26] | 47.33 / 77.66 / 58.81 / 41.66 / 95.63 | 87.67 / 76.03 / 79.91 / 65.53 / 97.21 | 66.45 / 54.21 / 59.71 / 42.56 / 87.57 |
| FC-Siam-Conc [26] | 60.88 / 73.58 / 66.63 / 49.95 / 97.04 | 88.31 / 79.86 / 83.45 / 71.43 / 98.43 | 59.67 / 65.71 / 62.54 / 45.50 / 86.63 |
| DTCDSCN [27] | 63.92 / 82.30 / 71.95 / 56.19 / 97.42 | 86.83 / 88.53 / 78.05 / 87.67 / 98.77 | 53.97 / 77.99 / 63.72 / 46.76 / 84.91 |
| STANet [44] | 79.37 / 85.50 / 82.32 / 69.95 / 98.52 | 83.81 / 91.00 / 87.26 / 77.40 / 98.66 | 67.71 / 61.68 / 64.56 / 47.66 / 88.49 |
| IFNet [28] | **96.91** / 73.19 / 83.40 / 71.52 / 98.83 | 94.02 / 82.93 / 88.13 / 78.77 / 98.87 | 53.94 / 67.86 / 60.10 / 42.96 / 87.83 |
| SNUNet [42] | 85.60 / 81.49 / 83.50 / 71.67 / 98.71 | 89.18 / 87.17 / 88.16 / 78.83 / 98.82 | 60.60 / 72.89 / 66.18 / 49.45 / 87.34 |
| BIT [31] | 86.64 / 81.48 / 83.98 / 72.39 / 98.75 | 89.24 / 89.37 / 89.31 / 80.68 / 98.92 | 68.36 / 70.18 / 69.26 / 52.97 / 89.41 |
| ChangeFormer [12] | 88.41 / 87.69 / 88.05 / 78.65 / 98.98 | 92.05 / 88.80 / 90.40 / 82.48 / 99.04 | 88.48 / 84.94 / 86.67 / 76.48 / 95.56 |
| VisionTwinNet | 94.53 / **88.79** / **91.57** / **84.45** / **99.35** | **92.93** / **90.87** / **91.88** / **84.04** / **99.19** | **92.56** / **91.90** / **92.22** / **85.57** / **97.36** |

* All values are reported in percentage (%). The highest scores are highlighted in bold.

make 7120/1024/2048 image pairs of training/validation/test, respectively.

DSIFN-CD [28]: This dataset, manually collected from Google Earth, consists of 6 large bi-temporal high-resolution images covering 6 cities in China, each with a size of 32507 × 15354. We do not follow its default dataset split and divide images of size 512 × 512 into non-overlapping patches of size 256 × 256. These patches are distributed into three parts to make 14400/1360/192 image pairs of training/validation/test, respectively.

WHU-CD [45]: This dataset, collected from WuHan University, consists of 5 large high-resolution images, covering over 220,000 buildings in Christchurch and New Zealand, each with a size of 32507 × 15354. We divide images of size 32507 × 15354 into non-overlapping patches of size 256 × 256. These patches are distributed into three parts to make 6096/762/762 image pairs of training/validation/test, respectively.

Low-LEVIR-CD: This dataset is a version of the LEVIR-CD dataset that has been manually processed for low-light conditions. It follows the configuration of LEVIR-CD.

LOL [46]: This dataset is a benchmark dataset used to address the real-world challenge of enhancing low-light images. It contains 500 pairs of images, each pair consisting of a low-light and unprocessed images, each with a resolution of size 400 × 600. The image pairs are clipped into 256 × 256.

### B. COMPARISON METHODS

In this section, we present the CD performance of the current state-of-the-art (SOTA) methods, as illustrated in TABLE 2.

FC-EF [26]: This method concatenates bi-temporal images and processes them through a single-stream convolutional network, based on the UNet architecture, to detect changes using an early fusion strategy.

FC-Siam-Diff [26]: This method is a feature-difference approach that uses a post-fusion strategy based on the FC-EF network. It extracts multiscale features from a twin convolutional network for bitemporal images. Algebraic operations are then applied to these features to obtain disparity features, which are used to detect changes.

FC-Siam-Conc [26]: This method is a feature-concatenation approach for change detection. It extracts multiscale features from twin convolutional networks. Unlike the FC-Siam-Diff method, the FC-Siam-Conc method concatenates these multiscale features in the channel dimension to perform change detection.

DTCDSCN [26]: This method is a dualtask constrained deep Siamese convolutional network (DTCDSCN). It consists of three sub-networks: a change detection network and two semantic segmentation networks (SSN). Additionally, a dual attention module (DAM) is introduced, along with an improved focal loss to address the issue of sample imbalance.

CDNet [47]: This method is based on an inverse convolutional network for change detection. Utilizing an early fusion strategy, it takes a pair of images as input and produces a pixel-level classification map highlighting structural changes.

STANet [44]: This method is a spatiotemporal attention neural network based on Siamese architecture. The network aggregates multi-scale feature maps from consecutive frames to leverage temporal consistency.

IFNet [28]: IFNet is a multi-scale feature concatenation method designed for change detection. It fuses multi-level deep features of bi-temporal images with image difference features. This fusion strategy allows IFNet to effectively detect changes between the two images.

SNUNet-CD [42]: This method combines Siamese network and NestedUNet for change detection in remote sensing images. It emphasizes the preservation of shallow-layer, high-resolution features often overlooked by other methods. The network uses the Ensemble Channel Attention Module (ECAM) for deep supervision, refining features for accurate classification.

Bit-CD [31]: This method efficiently models spatial-temporal contexts in remote sensing change detection. It translates bitemporal images into semantic tokens, refines them using transformers, and outperforms many state-of-the-art methods with fewer computational costs.

ChangeFormer [12]: Unlike traditional convolutional-based methods, ChangeFormer combines a hierarchically structured transformer encoder with a Multi-Layer Perception (MLP) decoder, enabling efficient capture of multi-scale long-range details for accurate change detection.

We employed the state-of-the-art methods, along with VisionTwinNet, to assess change detection performance on the WHU-CD, LEVIR-CD, and DSIFN-CD test sets. TABLE 2 presents a comprehensive performance comparison of VisionTwinNet with other state-of-the-art techniques on the WHU-CD, LEVIR-CD, and DSIFN-CD test sets. Quantitative results demonstrate that VisionTwinNet exhibits superior performance across these three CD datasets. On the WHU-CD dataset, the precision, recall, F1 score, and IoU are 94.53%, 88.79%, 91.57%, and 84.45%, respectively, surpassing the ChangeFormer model by 6.12%, 1.1%, 3.52%, and 5.8%. On the LEVIR-CD dataset, VisionTwinNet still outperform other models, achieving 92.93%, 90.87%, 91.88%, and 84.04% in terms of precision, recall, F1 score, and IoU. Given the high accuracy of models on this dataset, VisionTwinNet slightly exceeds ChangeForer by 0.88%, 2.07%, 1.48%, and 1.56%. On the DSIFN-CD dataset, our model showcases the best performance, with scores of 92.56%, 91.90%, 92.22%, and 85.57% respectively. Notably, our model's F1 score surpasses that of ChangeFormer by 5.55%. It's worth highlighting that our VisionTwinNet model demonstrates exceptional performance across all three datasets.

The visual comparison of various CD methods across the WHU-CD, LEVIR-CD, and DSIFN-CD datasets is depicted in Figure. 7. To highlight the differences in detection results, we employed red squares to underscore distinct areas of disparity. It is evident that our model yields superior results compared to others. VisionTwinNet performs exceptionally well in capturing small local changes, intricate regions of dense and subtle variations, and accurately differentiates between building boundaries and their shadows, as seen in Figure. 7(c), (d), and (e). This is attributable to the combination of LFCU and LR-Transformer, which equips the model with a robust capability to capture detailed features and long-range information. Moreover, our model demonstrates resilience against fine-grained changes in dense areas and variations in similar color regions, as illustrated in Figure. 7(e), (f), and (g). For instance, in Figure. 7(e), the VisionTwinNet model adeptly extracts features of narrow roads or small object changes, as indicated by the red box. Compared to other models, ours benefits from learning long-range context, resulting in fewer missed detections than models like BIT-CD. These results can be observed in Figure. 7(e) and (f). Additionally, our model, benefits from late fusion, effectively retains and captures detailed features from high-resolution satellite images. Distinctive contributions from different feature levels ensure smooth boundaries, precision, and apt handling of subtle changes. Notably, compared to scenarios like in Figure. 7(h) where EF-CF over-detects buildings and roads beyond their actual size, our model exhibits fewer false detections than other benchmark methods. Furthermore, our model is robust against seasonal and lighting changes that introduce noise, as well illustrated by TABLE 3.

**TABLE 3. Quantitative comparison results of different CD methods on Low-LEVIR-CD. All values are reported in percentage (%). Low-LEVIR-CD is a manually processed low-light LEVIR-CD dataset.**

| Method | Low-LEVIR-CD | | | | |
|---|---|---|---|---|---|
| | Precision | Recall | F1 | IoU | OA |
| ChangeFormer | 84.69 | 20.36 | 32.83 | 19.64 | 95.76 |
| CRDNet_base | 63.26 | 32.53 | 42.97 | 27.36 | 95.60 |
| VisionTwinNet | 79.31 | 71.66 | 75.29 | 60.36 | 97.31 |

**TABLE 4. Quantitative comparison results on LOL dataset in terms of different image quality methods. The highest scores are highlighted in bold.**

| Metrics | CRM | LIME | MF | Retinex-Net | NPE | GLAD | KinD++ | Ours |
|---|---|---|---|---|---|---|---|---|
| PSNR | 17.2033 | 16.7586 | 18.7916 | 16.7740 | 16.9697 | 19.7182 | 21.3003 | **22.9713** |
| SSIM | 0.6442 | 0.5644 | 0.6422 | 0.5594 | 0.5894 | 0.7035 | 0.8226 | **0.8867** |
| $LOE_{REF}$ | 926.1 | 1342.4 | 1042.1 | 2201.7 | 1643.1 | 1017.1 | 776.2 | **704.3** |
| NIQE | 7.6865 | 8.3777 | 8.877 | 8.8785 | 8.439 | 6.4755 | 3.8807 | **3.9936** |

TABLE 3 demonstrates the robustness of VisionTwinNet, when equipped with Gated EnhanceClearNet, in high-precision change detection of remote sensing bi-temporal images subject to seasonal and lighting condition variations. In the Low-LEVIR-CD dataset, the F1 score of VisionTwinNet is 42.46% higher than that of ChangeFormer, 32.32% higher than CDNet, and the IoU is 40.72% higher than ChangeFormer and 33% higher than CDNet. CDNet is a version of the VisionTwinNet algorithm that does not incorporate Gated EnhanceClearNet. Owing to the parallel and serial combination of Transformer and CNN, its detection performance in the Low-LEVIR-CD dataset, which is characterized by low illumination and significant image noise, remains superior to ChangeFormer.

TABLE 4 presents the quantitative comparison results among various image quality methods on the LOL dataset. To measure and compare the performance of different benchmark methods, we utilized four commonly used image quality metrics, namely PSNR, SSIM, LOE, and NIQE. PSNR is a frequently used metric to measure the quality of image or video compression, with higher values indicating better reconstructed image quality. SSIM, on the other hand, is a metric that accounts for image structure, brightness, and other details to provide a judgment that more closely aligns with human visual perception. It ranges from -1 to 1, with 1 indicating identical images. LOE is an index reliant on perceived luminance order [48]. NIQE is a no-reference image quality assessment method [49], where a lower score typically denotes superior quality, closely resembling natural images. Our EnhanceClearNet network significantly outperforms other benchmark methods. Using the ground truth images provided by the LOL dataset as reference, we computed the PSNR and SSIM scores, further underscoring that EnhanceClearNet considerably surpasses other benchmark

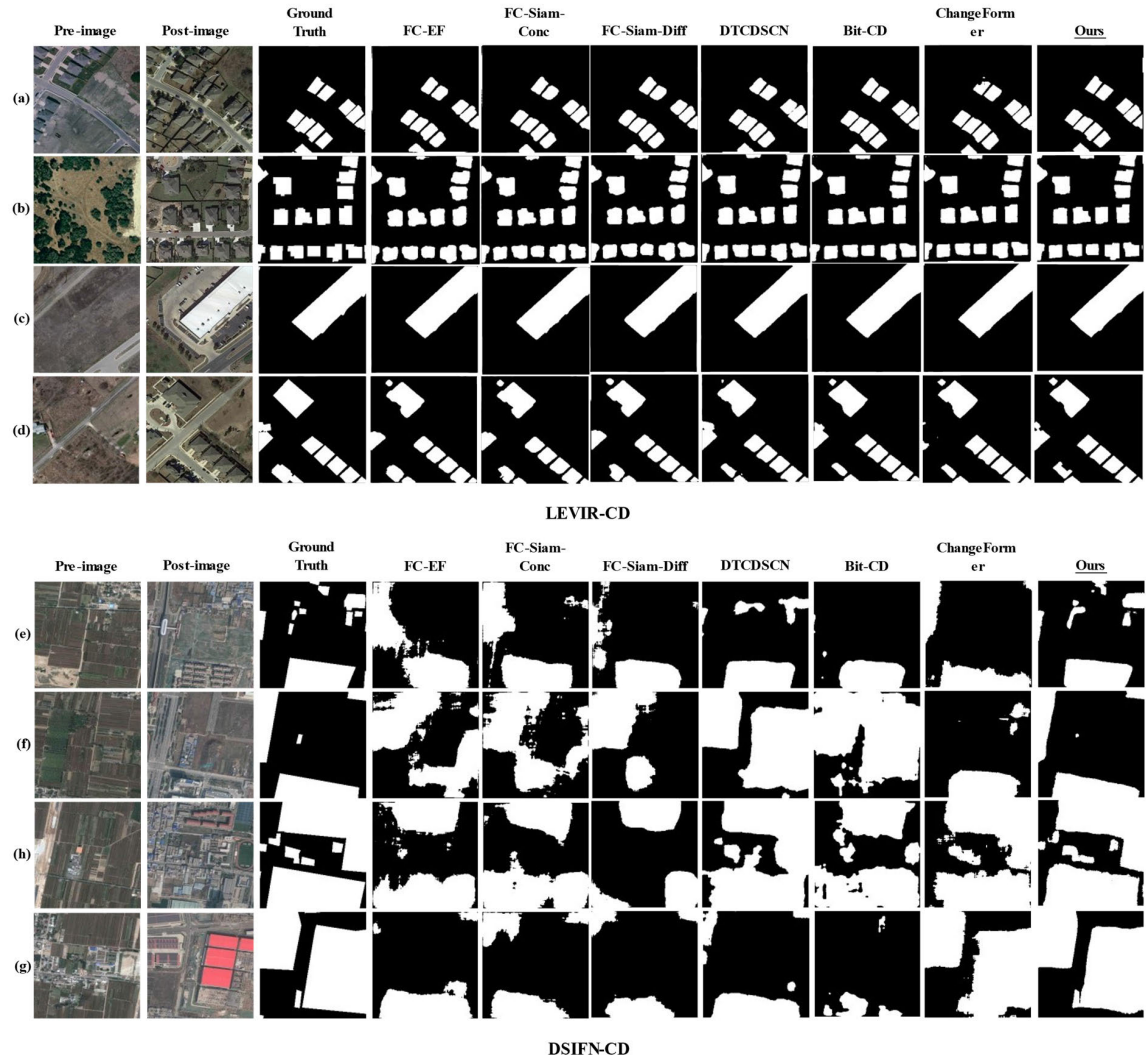


**FIGURE 7.** Visual comparison of different CD methods on the LEVIR-CD and DSIFN-CD test sets.

methods. Similarly, our model also surpasses other methods in LOE and NIQE metrics, exceeding the second-best by 71.9 and 0.1129 respectively. Moreover, EnhanceClearNet demonstrates stronger robustness and generalization capabilities, performing superiorly on high-precision bi-temporal remote sensing images like LEVIR-CD when compared to other methods.

### C. IMPLEMENTATION DETAILS

For our experiments, we utilized the TensorFlow framework on an Nvidia GTX 3090Ti GPU and an Intel Core i9-12700 5.0GHz CPU. The implementation details will be explained in two sections: the first emphasizes the Gated EnhanceClearNet for image enhancement, while the latter focuses on the Hybrid Light-Robust CDNet for change detection. Within the image enhancement domain, the Decomposer is trained with a batch size of 12 and processes image patches of dimensions $48 \times 48$. Moreover, the EnhanceNet and ClearNet are trained with a batch size of 4, handling image patches sized $256 \times 256$. In terms of change detection, our training scheme incorporates data augmentation strategies, including Gaussian blur and random flipping. Additionally, Cross-Entropy (CE) loss and adaptive moment estimation (AdamW) optimizer are used to train the model with a weight decay of 0.02. The learning rate starts at 0.00001 and decreases linearly to 0 over time. The model is trained for 300 epochs with a batch size of 16. To maintain a fair comparison with other methods, we have re-implemented all the change detection networks from publicly available codes, adhering to their default hyperparameters.

### D. EVALUATION METRICS

To fully evaluate the performance of the model in change detection, we've selected a set of metrics that are both widely recognized and particularly relevant to the nuances of the task. We report on precision (Pre.) and recall (Rec.) to capture

**TABLE 5.** Complexity and performance evaluation of different CD methods on LEVIR-CD dataset.*

| Methods | Flops(G) | Params(M) | F1(%) |
|---|---|---|---|
| FC-EF [26] | 2.34 | 1.31 | 79.36 |
| FC-Siam-Di [26] | 3.56 | 1.61 | 79.91 |
| FC-Siam-Conc [26] | 3.97 | 1.88 | 83.45 |
| DTCDSCN [27] | 27.13 | 11.57 | 78.05 |
| STANet [44] | 6.58 | 16.93 | 87.26 |
| IFNet [28] | 41.18 | 50.71 | 88.13 |
| SNUNet [42] | 27.44 | 12.03 | 88.16 |
| BIT [31] | 10.60 | 3.55 | 89.31 |
| ChangeFormer [12] | 41.02 | 106.22 | 90.40 |
| VisionTwinNet | 37.91 | 97.06 | 91.88 |

the model's accuracy and sensitivity in identifying changes. The overall accuracy (OA) offers a broad view of the model's general performance. The F1 score, acting as the harmonic mean of precision and recall, and the mean Intersection over Union (mIoU), which quantifies the average overlap between predicted and actual results, are central to our assessment, providing a nuanced understanding of the model's efficacy in delineating change categories. Detailed computations for these metrics are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{27}$$

$$mIoU = \frac{TP}{TP + FP + FN} \tag{28}$$

$$OA = \frac{TP + TN}{TP + FN + TN + FP} \tag{29}$$

### E. COMPLEXITY AND PERFORMANCE EVALUATION

We propose a novel change detection method. It ingeniously integrates the Transformer with convolutional neural networks for encoding, and innovatively introduces several useful modules such as Local Feature Capture Unit, Image Augmentation Mechanism, and Scale Adaptive Decoder Mechanism. Inevitably, these modules have a negative impact on model light-weighting. To control model complexity and maintain reasonable runtime, we introduced a lightweight self-attention mechanism and optimized the fusion strategy, reducing the layers of the model. As a result, the current parameter and computation load of the VisionTwinNet network are acceptable. While maintaining superior performance, VisionTwinNet operates with a shorter runtime compared to IFNet and ChangeFormer and boasts a smaller parameter count than ChangeFormer. In TABLE 5, we evaluate the model performance with two complexity indicators, which is the number of parameters and the number of floating-point operations (FLOPs). Despite implementing a lightweight self-attention module, optimizing fusion strategies, and incorporating gating mechanisms, our model still possesses the second-highest number of parameters compared to other methods due to the addition of numerous modules. However, its computational complexity remains acceptable within an academic context. Most crucially, the accuracy of our model has improved significantly.

### F. ABLATION STUDY

In the ablation study section of this paper, we discuss the impact and roles of different modules on performance. We further delve into the effects of various encoders and decoders on the performance, as well as the influence of different fusion strategies and datasets on the model's efficacy.

#### 1) IMPACT OF MODULES ON PERFORMANCE

In this paper, ablation experiments based on LEVIR-CD dataset is carried out to assess the contributions of different modules of our VisionTiwnNet to its overall performance. By removing or combing components of the model, we can discuss effectiveness and significance of each module in this network. As illustrated in TABLE 5, we present the quantitative results of the ablation study for various model components. Additionally, to provide a more intuitive understanding of the impact of these components, Figure. 8 offers a visual comparison. Within the figure, ''W/O'' is an abbreviation for ''without'', indicating the absence of the respective component.

The comparison between (2) and (4) in TABLE 6 confirms that the Gated EnhanceClearNet (GECN) Module has no significant effect on standard datasets, such as the LEVIR-CD dataset. As shown in TABLE 3, this module does enhance change detection (CD) performance and accelerates training for low-light datasets. From the visualization in Figure. 8, it can be observed that VisionTwinNet without the Gated

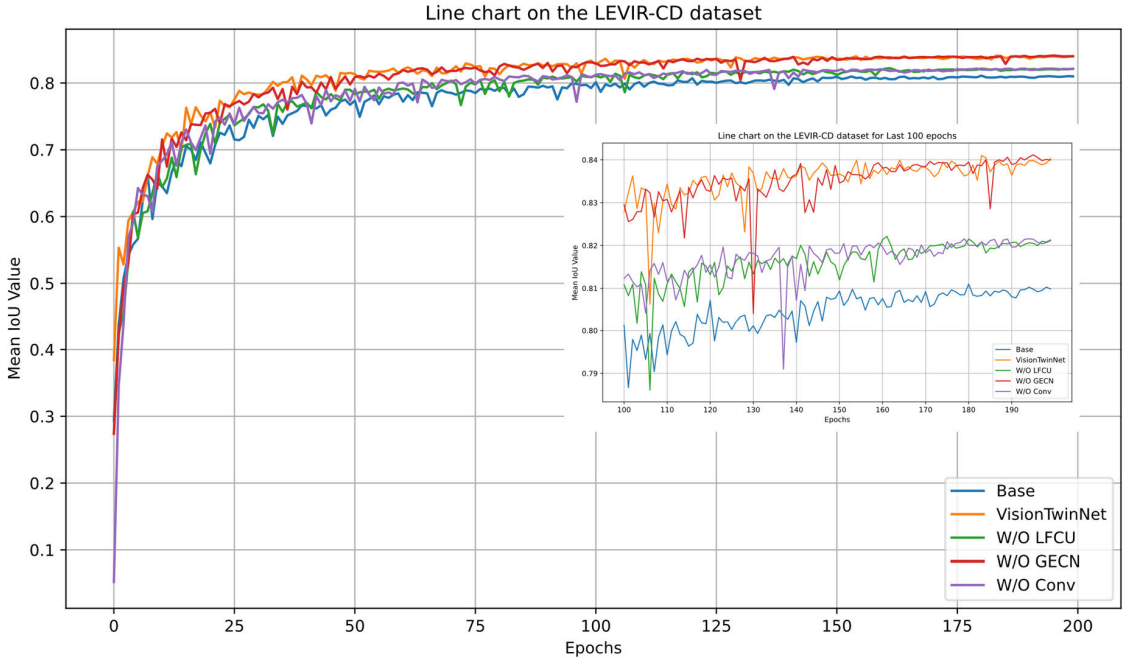


**FIGURE 8.** Accuracy comparison of ablation. (1)-(5) denote Base, VisionTwinNet, W/O LFCU, W/O GECN, W/O Conv, respectively.

**TABLE 6.** Analysis of ablation study on different modules.

| Method | LEVIR-CD | | | | |
|---|---|---|---|---|---|
| | Precision | Recall | F1 | IoU | OA |
| Base | 90.75 | 87.33 | 89.01 | 80.84 | 98.01 |
| W/O LFCU | 91.42 | 88.05 | 89.70 | 81.99 | 98.07 |
| W/O Conv | 91.62 | 87.88 | 89.71 | 82.01 | 98.07 |
| W/O GECN | 93.04 | 90.16 | 91.58 | 84.03 | 99.06 |
| VisionTwinNet | 92.93 | 90.87 | 91.88 | 84.04 | 99.19 |

* All values are reported in percentage (%).
* (1)-(5) denote Base, VisionTwinNet, W/O LFCU, W/O GECN, W/O Conv, respectively.

EnhanceClearNet Module shows a difference in accuracy only in the early stages of training compared to the complete VisionTwinNet. However, throughout the training process, the IoU accuracy curve remains consistent with the full VisionTwinNet. Moreover, the final F1 and IoU accuracies only differ by 0.3% and 0.01% respectively. This shows that the module does not affect the accuracy of our change detection algorithm under conventional scenarios. Its primary utility is evident in scenarios where the dataset image quality is suboptimal, highlighting its robust capability for change detection under adverse imaging conditions.

And the comparison between (2) and (5) in TABLE 6 validates that the high-level features captured by the Conv Block contribute significantly to the spatial information recovery of feature maps in VisionTwinNet, leading to enhanced change detection performance with F1 and IoU enhancing 2.17% and 2.03% respectively. This is further confirmed by the training accuracy curve shown in Figure. 8.

The comparison between (2) and (3) in TABLE 6 shows that VisionTwinNet with the Local Feature Capture Unit improves change detection performance. Specifically, F1 scores improved significantly from 89.70% to 91.88%, while IoU scores increased from 81.99% to 84.04%. As shown in Figure. 8, the Local Feature Capture Unit (LFCU) is adept at comprehensively capturing local features and contextual information to improve CD performance. Moreover, it also compensates for the common shortcoming of the Transformer architecture, which might overlook local details when capturing long-distance information. As a result, during training, the curve is smoother and reaches a gentler curvature more rapidly. Additionally, as illustrated in Figure. 9, it aids VisionTwinNet in recovering more hierarchical and detailed spatial information. From these figures, it can be observed that networks equipped with LFCU can predict changes more accurately, reducing the occurrence of erroneous change detections. In summary, this paper separately discusses the impacts of the IAM, LPCU, lightweight-attention, and Conv Block modules on overall performance.

### 2) IMPACT OF FUSION STRATEGIES ON PERFORMANCE

To optimize our experimental performance, we carried out a comparative analysis between our proposed fusion strategies: early fusion and late fusion, based on the LEVIR-CD dataset. The findings, as illustrated in TABLE 7, reveal that the early fusion approach underperforms. Employing early fusion tends to ignore intricate feature details, hindering the extraction of more profound insights. On the other hand, late fusion ensures that each data type or feature set is initially

**TABLE 7. Comparison of fusion strategies.***

| Method | WHU-CD<br>Pre. / Rec. / F1 / IoU / OA | LEVIR-CD<br>Pre. / Rec. / F1 / IoU / OA | DSIFN-CD<br>Pre. / Rec. / F1 / IoU / OA |
|---|---|---|---|
| Early fusion | 71.63 / 67.37 / 69.37 / 53.11 / 97.61 | 84.91 / 74.36 / 79.36 / 64.83 / 97.98 | 72.61 / 52.73 / 61.09 / 43.98 / 88.59 |
| Late fusion | 47.33 / 77.66 / 58.81 / 41.66 / 95.63 | 87.67 / 76.03 / 79.91 / 65.53 / 97.2 | 66.45 / 54.21 / 59.71 / 42.56 / 87.57 |

* All values are reported in percentage (%).

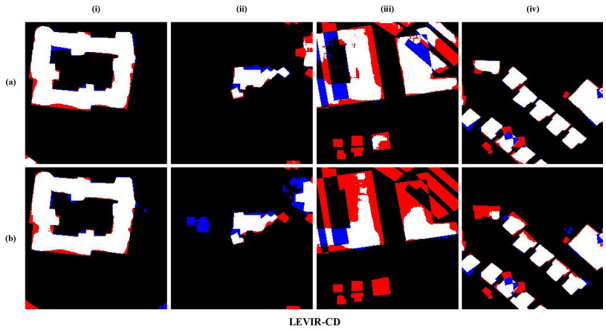


**FIGURE 9. Visualization results of VisionTwinNet and W/O LFCU on LEVIR-CD datasets. (a) and (b) denote VisionTwinNet and W/O LFCU, respectively. Red indicates FN, blue indicates FP.**

processed separately. This ensures the preservation of the unique attributes of each modality, preventing premature mixing or inadvertent loss of details. Thus, in this study, we opt to fuse the features extracted from various modules during the later stages, enabling the network to capture more detailed feature information.

## V. CONCLUSION

In this work, we introduced a novel two-stage approach, VisionTwinNet, that combines image enhancement and change detection for remote sensing images taken at different times. This not only addresses the limitations of using transformers in a brute-force manner in the computer vision domain and the high complexity of self-attention but also enhances the efficiency and flexibility of the model. It also ensures that the overall performance of the model is elevated without compromising visual quality. The proposed VisionTwinNet architecture utilizes both CNN and transformer to capture local and global information, enhancing the network's representation capability. Additionally, through the Lightweight Multi-Head Self-Attention mechanism, the model can process multiple distinct attention heads in parallel across different subspaces, which enhances the representation capability of the model. This, coupled with the innovative attention mechanism design that individually optimizes the key-value (k/v) dimensions, greatly enhances computational efficiency and reduces computational load. Notably, the VisionTwinNet structure incorporates an adaptive image enhancement strategy based on a gating mechanism. The core idea of this strategy is to automatically decide, based on the quality of the input image, whether image enhancement and degradation removal are required, significantly boosting the model's adaptability and performance for images of varying quality. Extensive experiments on three public datasets: LEVIR-CD, DSIFN-CD, and WHU-CD, demonstrate that our proposed VisionTwinNet performs well in terms of comprehensive performance. While this research has enhanced robustness to different illumination conditions and simultaneously improved the model's flexibility and performance, the real-time performance of the model remains a challenge. Although the computational cost of transformers is relatively low, their application in real-time change detection scenarios and handling vast change scenarios is still an area requiring further development. Our future work will aim to simplify the network structure and design more efficient architectures to refine remote sensing image change detection results.

## ACKNOWLEDGMENT

*(Tianao Chen and Aotian Chen are co-first authors.)*

## REFERENCES

[1] W. Li, M. Lu, and X. Chen, "Automatic change detection of urban land-cover based on SVM classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Milan, Italy, Jul. 2015, pp. 1686–1689.
[2] G. L. Angelici, N. A. Bryant, and S. Z. Friedman, "Techniques for land use change detection using Landsat imagery," in *Proc. Annu. Meeting-Amer. Soc. Photogramm.*, 1977, pp. 2–3.
[3] A. A. Abuelgasim, W. D. Ross, S. Gopal, and C. E. Woodcock, "Change detection using adaptive fuzzy neural networks," *Remote Sens. Environ.*, vol. 70, no. 2, pp. 208–223, Nov. 1999.
[4] J. Im and J. Jensen, "A change detection model based on neighborhood correlation image analysis and decision tree classification," *Remote Sens. Environ.*, vol. 99, no. 3, pp. 326–340, Nov. 2005.
[5] L. Auret and C. Aldrich, "Change point detection in time series data with random forests," *Control Eng. Pract.*, vol. 18, no. 8, pp. 990–1002, Aug. 2010.
[6] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Jan. 1995.
[7] W. Gu, Z. Lv, and M. Hao, "Change detection method for remote sensing images based on an improved Markov random field," *Multimedia Tools Appl.*, vol. 76, no. 17, pp. 17719–17734, Sep. 2015.
[8] H. Zhang, Y.-J. Hu, J.-W. Chen, and L.-W. Xiu, "Realization of translation group in optical design with deep nerual network under Eikonal-energy mapping," *Acta Phys. Sinica*, vol. 71, no. 13, 2022, Art. no. 134201.
[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, Oct. 2012, pp. 1106–1114.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.
[12] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Lumpur, Malaysia, Jul. 2022, pp. 207–210.

[13] L. Fazry, M. M. L. Ramadhan, and W. Jatmiko, "Change detection of high-resolution remote sensing images through adaptive focal modulation on hierarchical feature maps," *IEEE Access*, vol. 11, pp. 69072–69090, 2023.
[14] D. Lu, S. Cheng, L. Wang, and S. Song, "Multi-scale feature progressive fusion network for remote sensing image change detection," *Sci. Rep.*, vol. 12, no. 1, p. 11968, Jul. 2022.
[15] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603216.
[16] H. Zhang, Y. Wu, H. Song, and K. Zhang, "Gradient-guided temporal cross-attention transformer for high-performance remote sensing change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[17] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1971.
[18] V. Vapnik, "The support vector method of function estimation," in *Nonlinear Modeling*, J. A. K. Suykens and J. Vandewalle, Eds. Boston, MA, USA: Springer, 1998, pp. 55–85.
[19] A. Mohsenifar, A. Mohammadzadeh, A. Moghimi, and B. Salehi, "A novel unsupervised forest change detection method based on the integration of a multiresolution singular value decomposition fusion and an edge-aware Markov random field algorithm," *Int. J. Remote Sens.*, vol. 42, no. 24, pp. 9376–9404, Nov. 2021.
[20] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, Mar. 2013, pp. 3–4.
[21] T. Kasetkasem and P. K. Varshney, "An image change detection algorithm based on Markov random field models," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 8, pp. 1815–1823, Aug. 2002.
[22] L. Zhou, G. Cao, Y. Li, and Y. Shang, "Change detection based on conditional random field with region connection constraints in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3478–3488, Aug. 2016.
[23] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 565–571, May 2018.
[24] G. Troglio, M. Alberti, J. A. Benediksson, G. Moser, S. B. Serpico, and E. Stefánsson, "Unsupervised change-detection in retinal images by a multiple-classifier approach," in *Multiple Classifier Systems*. Cairo, Egypt: Springer, 2010, pp. 94–103.
[25] J. G. Vinholi, D. Silva, R. Machado, and M. I. Pettersson, "CNN-based change detection algorithm for wavelength-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[26] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 4063–4067.
[27] 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.
[28] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.
[29] K. Dwivedi, M. F. Bonner, R. M. Cichy, and G. Roig, "Unveiling functions of the visual cortex using task-specific deep neural networks," *PLOS Comput. Biol.*, vol. 17, no. 8, Aug. 2021, Art. no. e1009267.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proc. 16th Asian Conf. Comput. Vis.*, Macao, China, 2023, pp. 75–92.
[32] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514, doi: 10.1109/TGRS.2021.3095166.
[33] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, "A transformer-based Siamese network and an open optical dataset for semantic change detection of remote sensing images," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 1506–1525, Sep. 2022.
[34] J. Wang, H. Koizumi, and T. Kamiya, "Accuracy improvement of change detection based on color analysis," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. B7, pp. 357–361, Aug. 2012.
[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MCCAI)*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2015, pp. 234–241.
[36] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," 2011, *arXiv:1101.3354*.
[37] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.
[38] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1013–1037, Jan. 2021.
[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.
[41] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2. New Orleans, LA, USA, 2018, pp. 464–468.
[42] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[43] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[44] 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, p. 1662, May 2020.
[45] 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.
[46] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.
[47] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auto. Robots*, vol. 42, no. 7, pp. 1301–1322, May 2018.
[48] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.
[49] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

**TIANAO CHEN** received the B.S. degree in computer science from the University of Leeds, Leeds, U.K., in 2023, and the B.E. degree in computer science from Southwest Jiaotong University, Chengdu, China, in 2023. He is currently pursuing the M.S. degree with the Electrical and Computer Engineering Department, University of Michigan, Ann Arbor, USA.

His research interests include computer vision, deep learning, and machine learning.

**AOTIAN CHEN** received the B.S. degree in computer science from Lancaster University, Lancaster, U.K., in 2023, and the B.E. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2023. He is currently pursuing the M.S. degree with the Electrical and Computer Engineering Department, Georgia Institute of Technology, Georgia, USA.

His research interests include computer vision, deep learning, and remote sensing image processing and analysis.