CSCNet: A Cross-Scale Coordination Siamese Network for Building Change Detection

Yiyang Zhao, Xinyang Song, Jinjiang Li, and Yepeng Liu

***Abstract*—Remote sensing image change detection (CD) has witnessed remarkable performance improvements with the guidance of deep learning models, particularly convolutional neural networks and transformers. Current CD methods heavily rely on multilayered backbone structures, such as ResNet and Unet, for feature extraction. However, these approaches exhibit limitations in coordinating the utilization of local and global features across different scales. In this article, we introduce a novel cross-scale coordinated siamese (CSC) network to effectively integrate multiscale information. We introduce a cross-scale coordination module (CSCM) within the CSC network to coordinate internal features of the local branch with cross-scale information from adjacent branches, while simultaneously attending to both the local and global regions. Furthermore, to comprehensively capture contextual information, we propose a transformer aggregation module as a decoder to harmonize the output features of CSCM. We extensively evaluate our proposed CSC network on three datasets, namely, LEVIR-CD, WHU-CD, and GZ-CD. The results demonstrate that our CSC network outperforms other leading methods significantly in terms of F1-score and intersection over union evaluation metrics.**

***Index Terms*—Convolutional neural network (CNN), cross-scale coordinated, remote sensing change detection (CD), transformer.**

I. Introduction

CHANGE detection (CD) in remote sensing is a critical research area that utilizes remote sensing techniques to compare multiple temporal images of the earth's surface, aiming to detect changes in surface coverings. With the continuous growth of the global population and rapid urbanization, human activities drive constant changes in land use and land cover.

CD in remote sensing images can be successfully applied in various scenarios, including urban management [1], damage assessment [2], forest logging [3], environmental monitoring [4], and agricultural changes [5], among others. Owing to the diversity of application scenarios and target features, the CD task faces challenges arising from different data conditions and performances. The imaging conditions of multitemporal images are also difficult to ensure complete consistency, which further increases the difficulty of the CD task.

Early CD methods primarily employed pixel-based CD approaches, treating each pixel as the fundamental processing unit. They extracted change information by comparing pixel information differences between pre- and post-remote-sensing images [6]. With the widespread application of machine learning techniques, remote sensing image segmentation methods have gradually been introduced into CD tasks. Traditional clustering [7] or threshold-based methods [8] can generate binary CD images but often struggle with handling change information within images. Principal component analysis (PCA) [7], [9] and change vector analysis (CVA) [10] are long-used techniques in CD methods to enhance the processing of change information in images. These technologies have proven effective in a variety of CD applications over the years, demonstrating their enduring relevance in this field. Unlike traditional clustering and threshold-based methods, PCA and CVA enhance the accurate detection of change regions because they can better capture image features and change information. Furthermore, they can further extract and analyze change information by analyzing image change vectors and principal components, thereby improving the effectiveness and robustness of CD methods.

Convolutional-based CD methods have already achieved superior performance in CD tasks compared to traditional methods. However, as high-resolution remote sensing images continue to advance, there is a growing need for more precise target identification and CD methods to meet the demands of complex and clear images. Owing to the limitations of receptive fields (RFs) in pure convolutional methods, many researchers have been working on enhancing global information extraction for high-resolution images and have focused on more efficient context modeling methods to identify changes more accurately in regions of interest. To expand the RF, researchers have tried various methods, including stacking more convolutional layers [11], [12], [13], [14] or employing dilated convolutions [13]. These methods aim to capture a broader range of information to improve the accuracy of CD. With the emergence of attention mechanisms, researchers have begun to view them as a new tool for more effective context modeling to enhance the accuracy of CD. Attention-based CD methods [11], [12], [15], [16], [17] have to some extent improved the accuracy of CD, but they typically rely on the feature extraction backbone of convolutional layers. In recent years, the application of self-attention mechanisms in the field of CD has gradually increased. Unlike

Manuscript received 6 September 2023; revised 24 October 2023 and 9 November 2023; accepted 23 November 2023. Date of publication 30 November 2023; date of current version 14 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62202268, Grant 62002200, Grant 62272281, and Grant 61972235 and in part by the Shandong Natural Science Foundation of China under Grant ZR2023MF026 and Grant ZR2022MA076. *(Corresponding author: Yepeng Liu.)*

Yiyang Zhao and Xinyang Song are with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264005, China (e-mail: 450614766@qq.com; 1041824895@qq.com).

Jinjiang Li and Yepeng Liu are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: lijinjiang@gmail.com; liuyepengdream@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3337999

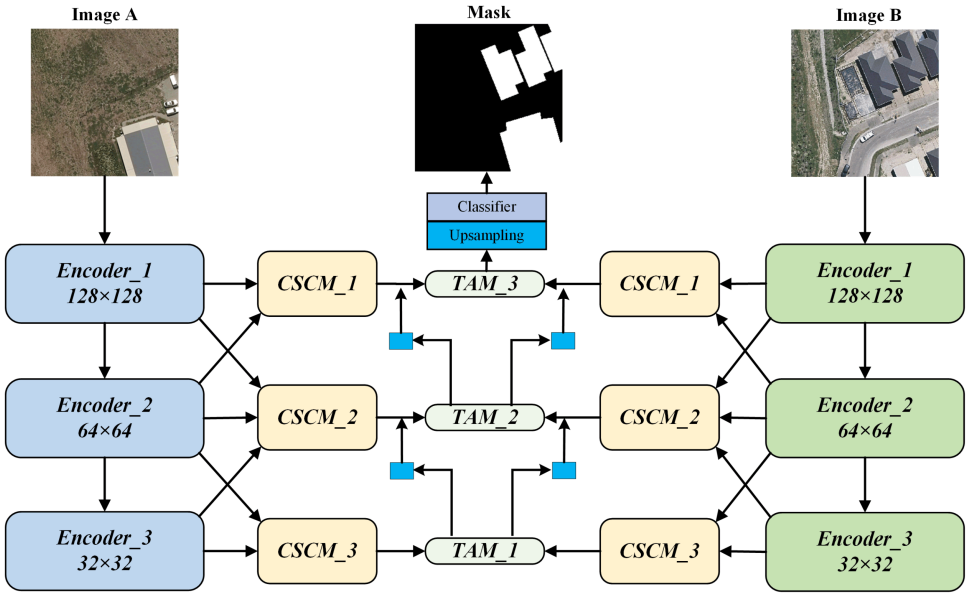

Fig. 1. Fundamental structure of the CSCNet comprises three key components: the encoder part, CSCMs, and the TAM. Initially, the encoder is responsible for extracting features at three different scales. Subsequently, these features pass through the CSCM to achieve cross-scale information coordination. This ensures that both local details and global features are fully utilized while considering the correlations between adjacent features. Next, in the TAM, the context modeling capabilities of the Transformer are employed to further optimize the fused features. Finally, a shallow CNN module is applied to generate feature masks. This entire process aids CSCNet in better capturing multiscale information within the images and enhances the performance of remote sensing CD.

previous channel or spatial-based attention mechanisms, self-attention mechanisms can better model relationships between pixels. However, the computational efficiency and complexity of self-attention mechanisms increase exponentially with the number of parameters, which is one of the challenges that current self-attention-based CD methods need to address.

In addition to its manifestation in the convolutional neural network (CNN), deep learning has also produced other powerful tools, such as generative adversarial networks [18], [19] and recurrent neural networks [20], [21], which have been applied to CD tasks. These deep-learning-based methods rely on their excellent modeling and feature extraction capabilities and do not require excessive design, making it possible to achieve the recognition of high-level information features and end-to-end CD of remote sensing data.

With the successful integration of Transformers into CD tasks [22] for remote sensing images, their exceptional contextual modeling capabilities have filled the gaps left by traditional convolutional and attention mechanisms in capturing long-range dense relationships. Remarkably improved results can be achieved even when using shallow CNN backbones. Furthermore, the CD field has witnessed the emergence of numerous Transformer-based methods [23], [24], [25], demonstrating the significant achievements of Transformers in the CD domain.

While many previous methods for CD in remote sensing images have attempted to optimize differential features by leveraging multiscale information to achieve more comprehensive feature representation, they have a limitation in how they interact with multiscale information, making it challenging to effectively relate information from different scales. Therefore, in this article, we introduce a cross-scale coordinated siamese network (CSCNet), as shown in Fig. 1. In the CSCNet, the cross-scale coordination module (CSCM) is applied to three stages of the encoder. In each CSCM, we parallelize convolution layers with different kernel sizes to capture both local and global content within a single feature level while relating features from different levels to each other. This design helps better capture the correlation information between features of different scales, which is particularly beneficial for optimizing details and determining the location of CD targets. In the decoder part, we employ the transformer aggregation module (TAM) to extend the RF of features, allowing for the capture of richer contextual information. These changes and improvements are expected to enhance the performance of CD methods in multiscale scenarios, better meeting the processing requirements of complex image data.

Our contributions can be summarized in the following points.

1) We improved the traditional codec architecture by introducing the CSCNet. We redesigned the ResNet structure and adopted depthwise overparameterized convolutional layer (DO-Conv) technology as the encoder for our network. This improvement not only surpasses the original ResNet in performance but also significantly reduces computational complexity. The DO-Conv [26] technique combines deep separable convolution and traditional convolution, which can improve network performance without increasing the amount of network inference calculation, and has a faster convergence speed than traditional convolution. By adopting DO-Conv technology, we aim to address some of the limitations of the existing work, such as improving the convergence speed and performance of the network, while reducing computational complexity, making our network more efficient and reliable when handling remote sensing image CD tasks.

2) We introduced the CSCM to facilitate the interaction of information within features and across multiple scales. This aids in capturing local details and global information, complementing each other through the decoder to achieve cross-level contextual information exchange.
3) We introduced the Transformer structure into the decoder, constructing the TAM. Simultaneously, we incorporated multiscale information from the CSCM into various decoder stages to capture multiscale contextual information.

The rest of this article is organized as follows. Section II provides background information and a review of traditional CD methods and deep-learning-based CD methods. In Section III, we delve into the detailed design and key components of our approach. Section IV summarizes and analyzes the extensive experimental results. Finally, Section V concludes this article.

II. Related Work

A. CNN-Based CD Method

CNNs have gained widespread application in various fields due to their powerful feature representation capabilities, including CD tasks in remote sensing imagery. In CD tasks, fully convolutional neural network (FCN) methods [27] have made significant advancements. However, traditional CNN methods are constrained by their limited local RFs, making it challenging to capture global information and restricting long-term modeling capabilities. To overcome this limitation, researchers, such as Song et al. [28], proposed the use of dilated convolutions and deformable convolutions as replacements for traditional convolutions to increase the RF and enhance context modeling capabilities.

Furthermore, attention mechanisms, as a crucial feature of CNNs, have been extensively applied in CD tasks. For instance, Li and Huo [29] improved the handling of feature difference maps using attention mechanisms. With the continuous development of attention mechanisms, innovative CD methods have emerged. These methods include Fang et al.'s [30] Unet++-based model, which uses channel attention to fuse information from different scales.

To further enhance performance, some studies have started to employ dense connections to fuse multiscale features or [30] introduce deep supervision methods [16], rather than relying solely on attention mechanisms. However, it is worth noting that the fixed local RFs in the FCN framework limit its long-term modeling capabilities. Therefore, there is a need to use larger CNN backbones and introduce more attention modules in deep convolutional stages, although this also increases the overall network complexity.

In this article, our objective is to leverage the proposed CSCM to its full extent for integrating multiscale contextual information. In addition, we introduce the Transformer into the decoder to extend the feature's RF, enabling the capture of richer contextual information. This is done to overcome the limitations imposed by the local RF.

B. Transformer-Based CD Method

In the field of computer vision, the Transformer has demonstrated significant potential and unique advantages compared to traditional attention mechanisms. Transformers utilize nonlocal attention mechanisms, which enable better modeling of global feature correlations, allowing for long-range dependencies between image pixels to be established more effectively. Currently, methods based on Transformers can be broadly categorized into two classes: pure Transformer methods [23], [24], which employ Transformers as the encoder, sharing parameters across multiple layers to capture long-range dependencies but incurring higher computational costs, and CNN-Transformer methods [5], [22], [25], [31], which optimize contextual information on top of the CNN, delivering improved performance while maintaining lower computational overhead. The work of Chen et al. [22] pioneered the use of Transformer and achieved excellent results when applied to shallow ResNet models, highlighting the superiority of Transformers in spatiotemporal modeling. Furthermore, some methods combine the Transformer with the CNN to enhance the RF for improved performance. For instance, Feng et al. [31] made full use of both the CNN and the Transformer to extract local and global features. Liu et al. [32] used ConvNets as the foundation to extract multiscale features from raw image pairs and effectively model contextual information in bitemporal images using attention and transformer modules. Transformer-based CD methods leverage Transformer's strengths in global feature modeling and long-range dependencies, leading to significant improvements in accuracy compared to traditional approaches. Therefore, Transformer-based methods hold great promise in the field of remote sensing CD.

III. Methodology

In this article, we propose a model called the CSCNet, and its overall architecture is depicted in Fig. 1. Our model follows an encoder–decoder structure.

1) *Encoder stage:* Bitemporal images, A and B, are passed through a modified ResNet34 network with DO-Conv. This results in three feature maps: $F_1^{E\_i}$, $F_2^{E\_i}$, and $F_3^{E\_i}, i = \{1, 2\}$, for each image.
2) *Cross-scale coordination modules:* To address the multiscale nature of our data, we use the CSCM. This module takes the aforementioned feature maps and generates new coordinated feature maps F1new, F2new, and F3new.
3) *Feature fusion:* This step combines the newly generated feature maps using the CAT operation to provide a consolidated representation. Feature maps are further refined using the TAM, which aggregates information across the scales to generate $F^{E\_1\_\text{new}}$, $F^{E\_1\_\text{new}}$, and $F^{E\_1\_\text{new}}$ after processing.
4) *Head CNN:* The final stage involves processing the TAM output through a head CNN to produce the predicted change mask, M.

These improvements made to the encoder introduce the traditional ResNet34 network with DO-Conv to slightly enhance

Algorithm 1: Implementation Process of Our CSCNet Model.

Input: A, B $(bitemporal\ image)$
Output: $M(a\ prediction\ change\ mask)$
// $step1 : Encoder\ stage$
$\mathbf{F_1^{E\_1}}, \mathbf{F_1^{E\_2}}, \mathbf{F_1^{E\_3}} = ResNet\_DOC(\mathbf{A})$
$\mathbf{F_2^{E\_1}}, \mathbf{F_2^{E\_2}}, \mathbf{F_2^{E\_3}} = ResNet\_DOC(\mathbf{B})$
// $step2 : Cross-scale\ coordination\ module$
for $i = \{1, 2\}$ **do**
  $\mathbf{F_i^{E\_1\_new}} = CSCM\_1(\mathbf{F_i^{E\_1}}, \mathbf{F_i^{E\_2}})$
  $\mathbf{F_i^{E\_2\_new}} = CSCM\_2(\mathbf{F_i^{E\_1}}, \mathbf{F_i^{E\_2}}, \mathbf{F_i^{E\_3}})$
  $\mathbf{F_i^{E\_3\_new}} = CSCM\_3(\mathbf{F_i^{E\_2}}, \mathbf{F_i^{E\_3}})$
end
// $step3 : Feature\ fusion$
$\mathbf{F^{E\_1\_new}} = CAT(\mathbf{F_1^{E\_1\_new}}, \mathbf{F_2^{E\_1\_new}})$
$\mathbf{F^{E\_2\_new}} = CAT(\mathbf{F_1^{E\_2\_new}}, \mathbf{F_2^{E\_2\_new}})$
$\mathbf{F^{E\_3\_new}} = CAT(\mathbf{F_1^{E\_3\_new}}, \mathbf{F_2^{E\_3\_new}})$
// $step4 : Transformer\ aggregation\ module$
$\mathbf{F_3} = TAM\_3(\mathbf{F^{E\_3\_new}})$
$\mathbf{F_2} = TAM\_2(\mathbf{F^{E\_2\_new}}, \mathbf{F_3})$
$\mathbf{F_1} = TAM\_1(\mathbf{F^{E\_1\_new}}, \mathbf{F_2})$
// $step5 : Head\ CNN$
$\mathbf{M} = Head(\mathbf{F_1})$

performance while significantly reducing computational complexity. The CSCMs serve to integrate multiscale feature information, and the decoder predominantly leverages the TAM. For a comprehensive overview of our method's inference process, we have provided specifics in Algorithm 1.

A. Encoder Module

In the current deep learning landscape, CNNs are among the most commonly used models. Among them, ResNet, as a classic deep CNN, has found extensive applications in various domains, including image classification and object detection. However, the traditional ResNet may encounter bottlenecks when dealing with complex images, leading to a decrease in performance, and its performance gradually loses its advantage concerning computational complexity. To address this issue, we have employed DO-Conv [26] to enhance the convolutional layers in ResNet, aiming to improve the model's flexibility, feature representation capacity, and reduce computational complexity.

DO-Conv is an improvement method for CNNs that combines the characteristics of depthwise separable convolution and overparameterized convolution. Traditional convolution applies the same convolutional kernel to all the channels when processing images, while depthwise separable convolution divides the convolution operation into depthwise convolution and pointwise convolution, allowing each channel to use different convolutional kernels, thus better extracting features. Overparameterized convolution, on the other hand, increases the number of convolutional kernels while keeping the kernel size unchanged to enhance the model's flexibility and expressive power. These improvements enable our model to better address challenges when dealing with complex images, thereby enhancing its performance.

DO-Conv has significant advantages over traditional CNNs as it can enhance the model's performance and accuracy while reducing computational complexity. By applying DO-Conv to ResNet, we have increased the model's flexibility and feature representation capabilities, enabling it to better handle complex images and improve the performance of CD tasks.

In this article, we used the first three stages of ResNet as the encoder, progressively reducing the feature size to 1/2, 1/4, and 1/8 of the original features. We also proposed two versions with different channel configurations: CSC_S with channel numbers of 64, 128, and 256, and CSC_L with channel numbers of 128, 256, and 512. In the actual implementation, we made corresponding adjustments to the subsequent modules based on the different channel numbers to ensure the effectiveness and performance improvement of the entire method. These improvements enable our model to better adapt to various task requirements and achieve performance optimization.

B. Cross-Scale Coordination Module

Contextual information plays a crucial role in processing high-resolution remote sensing images, and this information is not only present within single-scale features but also exhibits diverse characteristics at different scales. Within the same feature hierarchy, we first process single-scale features by using convolutional layers with different convolution kernels to capture both internal details and global features. Subsequently, we employ bidirectional attention mechanisms to extract feature representations that aggregate information from different convolutional kernel features. Between feature hierarchies at adjacent levels, we utilize spatial scale information for cross-scale feature extraction to obtain cross-level contextual complementary information. This approach allows high-level features to provide fine-grained information, while low-level features offer global cues, thereby coordinating the extraction of holistic features. Cross-scale interaction and aggregation are highly effective for refining details and determining target locations.

The CSCM is a crucial component bridging the encoder and the decoder, and its detailed operations are illustrated in Fig. 2. Taking into consideration the number of layers and different scales in the encoder, we have designed three branches of CSCM: the high-level feature branch, the low-level feature branch, and the local branch. It is pivotal to mention that the CSCMs processing Images A and B are identical in structure and function. CSCM_1 comprises the local branch and the low-level feature branch, CSCM_2 includes the local branch, the high-level feature branch, and the low-level feature branch, and CSCM_3 consists of the local branch and the high-level feature branch. We represent the CSCM operation as "f," and the description of the entire module is as follows:

$$F_{\text{CSCM}}^{i} = \begin{cases} f\left(F_E^i, F_E^{i+1}\right), & i = 1 \\ f\left(F_E^{i-1}, F_E^i, F_E^{i+1}\right), & i = 2 \\ f\left(F_E^{i-1}, F_E^i\right), & i = 3 \end{cases} \tag{1}$$

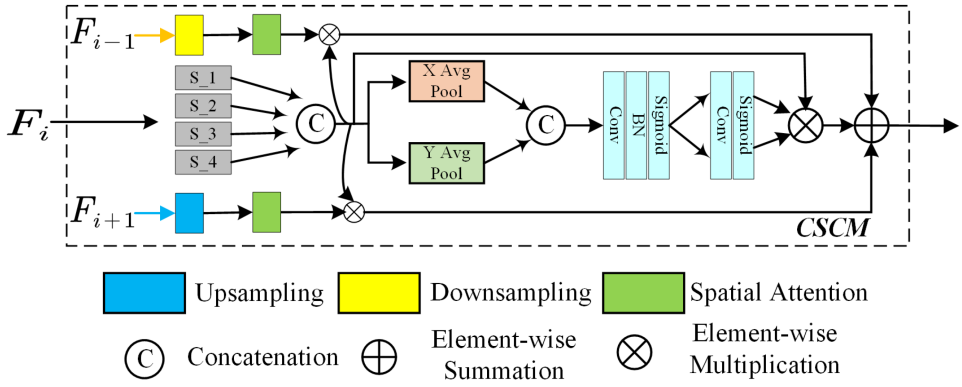

Fig. 2. CSCM simultaneously utilizes high-level, low-level, and local features for feature extraction and coordination. First, low-level features extract low-level detailed information through local convolution operations to form low-level feature maps. Second, high-level features extract high-level semantic information through global convolution operations to form high-level feature maps. Finally, local features are used to transfer and fuse contextual information, and low- and high-level feature maps are fused to form multiscale feature maps. This branch is designed to achieve the coordination and fusion of features at different scales to improve the stability and robustness of feature extraction.

where $F^i_{\text{CSCM}} \in \mathbb{R}^{C_i \times H_i \times W_i}$ represents the features output by each stage of the CSCM, and F_{i-1}, F_i, and F_{i-1} represents high-level features, current features, and low-level features, respectively.

1) Internal Feature Coordination: When dealing with remote sensing images, it is often necessary to capture both local and global feature information across multiple scales. To achieve this objective, we employ an extended convolution method based on DO-Conv, similar to what was mentioned earlier [33]. This convolution method enlarges the RF by using different kernel sizes and dilation rates to comprehensively capture contextual information. These contextual pieces of information will be decomposed into two 1-D feature encodings, followed by an attention operation. We integrate the features in both the spatial directions to simultaneously capture long-range dependencies while preserving precise positional information.

For the local branch features F^i_E, we first perform multiscale operations using four dilated convolutions, defined as follows:

$$F^{i,j}_{\text{DOC}} = \text{DOC}(F^i_E; K^{i,j}_{3\times3}; r^j), \quad j \in \{1,2,3,4\} \tag{2}$$

where $F^{i,j}_{\text{DOC}} \in \mathbb{R}^{C_i \times H_i \times W_i}$ represents the output feature, i represents the stage of the encoder, and j represents the level of the dilated convolution. The DOC operation includes the expanded convolution of DO-Conv and the batch normalization (BN) and ReLU activation functions. $K^{i,j}_{3\times3}$ represents the kernel size of 3×3, and r^j represents the expansion rate. Then, we perform the concatenation operation of the four features in the channel dimension, so that the output features have rich context clues and continue to process with the DOC of $r = 1$, which is defined as follows:

$$F^i_C = \text{DOC}_1(\text{Concat}(F^{i,1}_{\text{DOC}}, F^{i,2}_{\text{DOC}}, F^{i,3}_{\text{DOC}}, F^{i,4}_{\text{DOC}}); K^i_{3\times3}). \tag{3}$$

However, the concatenated feature information may contain redundant information. Therefore, we decompose the features into two 1-D feature maps and perform attention in different spatial dimensions. We adaptively adjust the feature weights of each position based on the spatial coordinate information of each position in the feature map, thereby allowing the model to focus on different positions in a more adaptive manner to purify the features F^i_C.

To begin with, we use global pooling to transform it into a pair of 1-D feature encodings. Specifically, for a given input $F^i_C \in \mathbb{R}^{C_i \times H_i \times W_i}$, we first use pooling kernels of size $(H_i, 1)$ or $(1, W_i)$ to encode each channel along the horizontal and vertical coordinates, respectively. Therefore, the output of channel c with height h can be represented as

$$z^h_c(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i). \tag{4}$$

Similarly, the output of the cth channel of width w can be written as

$$z^w_c(w) = \frac{1}{H} \sum_{0 \le i < H} x_c(j, w). \tag{5}$$

These two transformations can be captured and preserved by the attention module, which captures long-range dependencies and precise spatial information along one spatial dimension. This helps the network to more accurately locate the target of interest.

Next, the two parts of transformed information are concatenated, and a convolutional transformation function is applied to generate an intermediate feature map $F \in \mathbb{R}^{C/r \times (H+W)}$ with spatial information in the horizontal and vertical directions, where r represents the reduction ratio. Then, the feature F is split into two separate tensors $F^h \in \mathbb{R}^{C/r \times H}$ and $F^w \in \mathbb{R}^{C/r \times H}$, which are transformed by two 1×1 convolutions f_h and f_w to have the same number of channels as the input. The formula is expressed as

$$g^h = \sigma\left(F_h\left(f^h\right)\right) g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{6}$$

where σ represents the sigmoid function. g^h and g^w will be used as attention weights to generate the final output $F^{i,j}_{\text{loc}}$, defined as follows:

$$y^i_{\text{loc}}(i,j) = F^i_c(x,y) \times g^h_c(x) \times g^w_c(y). \tag{7}$$

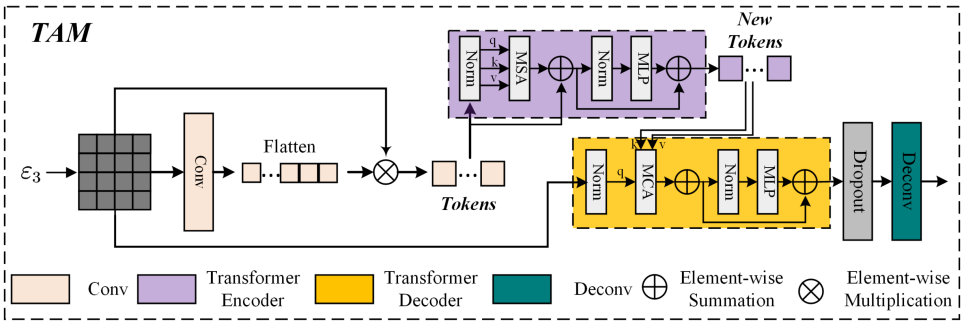

Fig. 3. TAM is a module composed of multiple Transformer layers; each layer contains MSA and feedforward network (FFN) sublayers. Input features are transformed through an MSA sublayer to generate a spatial attention map, and a weighted average is used to generate a new feature representation. The FFN sublayer uses two fully connected layers and a ReLU activation function to process the input features, and then, the outputs of MSA and FFN are spliced together to form the output features of this layer. Between each Transformer layer, the input features are skip-connected via residual connections. Finally, the output features are processed through the MCA sublayer, and the attention weights are obtained by calculating the dot product between Q and K and using the softmax function and then used to weight V to generate the final output features. The final output features are transformed through a fully connected layer to generate the final output.

2) Adjacent Feature Coordination: The operations on adjacent branches are much simpler compared to those on the local branch. We only use spatial attention to coordinate the information of other scales with the local features. In the first stage of operations, the coordination of features is performed between the low-level branch and the local branch, which can be expressed as

$$F_{ad\_u}^{1} = SA\left(\text{Up}\left(F_{E}^{2}\right)\right) \otimes F_{C}^{1} \tag{8}$$

where $\text{Up}(\cdot)$ implements a $2\times$ upsampling method through a bilinear interpolation method. Through this branch structure, the location information of the target can be brought into the feature information F_C^i of the current branch.

Similarly, in the operation of the third stage, the advanced branch and the local branch are used for feature coordination, expressed as

$$F_{ad\_d}^{1} = SA\left(\text{Down}\left(F_{E}^{2}\right)\right) \otimes F_{C}^{3} \tag{9}$$

where $\text{Down}(\cdot)$ realizes $2\times$ downsampling through maximum pooling. Through this branch structure, more detailed alignment information can be brought into the feature information of the current branch.

In the second stage, we utilize the high-level features, low-level features, and local features simultaneously to perform feature coordination. This stage includes the operations of the other two stages, and spatial attention is applied to both the high- and low-level features. It can be represented as

$$\begin{aligned} F_{ad\_u}^{2} &= SA\left(\text{Up}\left(F_{E}^{3}\right)\right) \otimes F_{C}^{2} \\ F_{ad\_d}^{2} &= SA\left(\text{Down}\left(F_{E}^{1}\right)\right) \otimes F_{C}^{2}. \end{aligned} \tag{10}$$

3) Branch Integration: After the effective coordination through internal and external feature fusion, we integrate the output features of all the branches with the original features. It can be represented as

$$F_{\text{CSCM}}^{i} = \begin{cases} F_{\text{loc}}^{i} \oplus F_{ad\_d}^{i} \oplus F_{E}^{i}, & i = 1 \\ F_{\text{loc}}^{i} \oplus \left(F_{ad\_u}^{i} \oplus F_{ad\_d}^{i}\right) \oplus F_{E}^{i}, & i = 2 \\ F_{\text{loc}}^{i} \oplus F_{ad\_u}^{i} \oplus F_{E}^{i}, & i = 3 \end{cases} \tag{11}$$

where $\oplus$ denotes elementwise summation. This cross-scale coordination of different levels of F_E^i can fully fuse global information and local detail information, greatly improving the stability and robustness of feature extraction.

C. Transformer Aggregation Module

The TAM is the basic unit of the decoder, as shown in Fig. 3. The feature information processed by the CSCM is first fed through a Transformer in the TAM to extract higher level contextual information, and then, the current CSCM feature information is aggregated with the TAM feature information from the previous step to generate a new input F_A^i for feature inference in the current step. We define the TAM process as $T(\cdot)$, and its formula is as follows:

$$\begin{aligned} F_{\text{TAM}}^{i} &= \begin{cases} T\left(F_{A}^{i}\right), & i = 1, 2 \\ T\left(F_{\text{CSCM}}^{i}\right), & i = 3 \end{cases} \\ F_{A}^{i} &= \text{Concat}\left(F_{\text{CSCM}}^{i}, \text{Deconv}\left(F_{\text{TAM}}^{i-1}\right)\right) \end{aligned} \tag{12}$$

where F_{TAM}^i is the output of the ith layer TAM and the input of the aggregation, and $\text{Deconv}(\cdot)$ means that the image is enlarged to a deconvolution layer with BN and ReLU of the same size.

The feature $F_A^i \in \mathbb{R}^{C_i \times H_i \times W_i}$ from the CSCM is first subjected to pointwise convolution for each pixel, followed by softmax computation for spatial attention mapping, and finally, the weighted average of the F_A^i pixel values is calculated. The input feature information is transformed into tokens $T \in \mathbb{R}^{L \times C}$ containing high-level concepts, where L represents the vocabulary size of the tokens, and is defined as follows:

$$T = \left(\sigma\left(\phi\left(F_{A}^{i}; W\right)\right)\right)^{T} F_{A}^{i}, \quad i \in \{1, 2\} \tag{13}$$

where $\phi(\cdot)$ represents the pointwise convolution of the learnable kernel $W \in \mathbb{R}^{C \times L}$, and $\sigma(\cdot)$ normalizes for each semantic group.

The obtained tokens T are passed into an N_E-layered Transformer composed of normalization, multihead self-attention (MSA), and multilayer perceptron (MLP). Here, MSA is the

classic operation in the Transformer. After these operations, a new set of tokens T^{new} is obtained.

The optimized tokens T^{new} will be projected back into the pixel space of the input feature F_A^i to obtain a better pixel-level feature representation enriched with contextual information. Compared to the Transformer, the decoding part has been modified by replacing MSA with multihead cross attention (MCA). In MCA, the input feature F_A^i serves as the query, while T^{new} serves as the key and value, which avoids excessive computations caused by dense relationships between pixels. This can be represented as

$$\text{MCA}\left(F_i^A, T^{\text{new}}\right) = \text{Concat}\left(h_1, h_2, \ldots, h_n\right) W^0 \tag{14}$$

where $W^0 \in \mathbb{R}^{hd \times C}$ is a linear projection matrix and n is the number of attention heads.

D. Other Network Details

1) Prediction Head: The final feature map we upsample by bilinear interpolation to double its size to generate a 256 × 256 feature map equal to the input size . Finally, two 3 × 3 convolution operations are performed to convert the channels into a variation mask of size $\mathbb{R}^{2\times H_0 \times W_0}$.

2) Loss Function: In the CD method, its essence is similar to the binary classification task of semantic segmentation, so we optimize the model parameters by minimizing the cross-entropy loss. The loss function is defined as follows:

$$L = \frac{1}{H_0 \times W_0} \sum_{h=1,w=1}^{H,W} l\left(P_{hw}, Y_{hw}\right) \tag{15}$$

where $l(P_{hw}, y) = -\log(P_{hwy})$ is the cross-entropy loss, and P_{hw} and Y_{hw} are the labels for the pixel at location (h, w).

IV. Experiments

A. Datasets

We use three high-resolution building CD datasets, namely, LEVIR-CD, WHU-CD, and GZ-CD.

1) LEVIR-CD [11] consists of 637 pairs of Google Earth images with a high resolution of 0.5 m/pixel, and each image has a size of 1024×1024 pixels. Our focus is on building-related changes, and we consider multiple types of buildings. To conduct experiments, we crop the LEVIR-CD dataset into nonoverlapping image patches of size 256×256 and divide them into training, validation, and test sets, which contain 7120, 1024, and 2048 image patches, respectively.
2) WHU-CD [34] is a large public CD dataset consisting of a pair of high-resolution images (0.075 m/pixel) with a size of 32 507×15 354 pixels. In the experiments, we cropped the dataset into 256×256 samples for training, validation, and testing, and the dataset size is 6096, 762, and 760, respectively.
3) GZ-CD [35] is a dataset consisting of 19 pairs of high-resolution (0.55 m/pixel) Google Earth images, including 19 pairs of seasonal change images covering urban changes in the suburbs of Guangzhou, China, over the past decade. The dataset mainly focuses on changes related to buildings, with image sizes ranging from 1006×1168 to 4936×5224. For experiments, we cropped the images into nonoverlapping 256×256 blocks and set the sizes of the training/validation/testing datasets to 2834/400/325, respectively.

B. Experimental Setup

1) Evaluation Metrics: In order to better assess the effectiveness of our method, we employ five metrics for the evaluation of various approaches: precision, recall, F1-score, intersection over union (IoU), and overall accuracy (OA). Among these, F1 and IoU are the primary evaluation metrics we focus on, where larger values indicate better model performance. Specifically, precision represents the ratio of true positive samples among those predicted as positive, recall represents the ratio of correctly predicted positive samples among the actual positives, and F1-score is the weighted harmonic mean of precision and recall. IoU represents the ratio between the intersection and union of predicted results and ground truth, while OA is the proportion of correctly classified samples. The expressions for these metrics are as follows:

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{F1}-\text{score} &= 2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \\
\text{Intersection over Union(IoU)} &= \frac{\text{TP}}{\text{TP}+\text{FN}+\text{FP}} \\
\text{Overall Accuracy(OA)} &= \frac{\text{TP}+\text{PN}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}}
\end{aligned} \tag{16}$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative, respectively.

2) Implementation Details: Our deep learning model was implemented using the PyTorch framework and trained in an environment running the Ubuntu operating system, with acceleration provided by a TITAN RTX GPU. We employed stochastic gradient descent with momentum as the optimizer, setting the momentum parameter to 0.99 and the weight decay to 0.0005. For all three datasets, we used a uniform learning rate of 0.01. During training, we performed performance validation using the validation set at the end of each training epoch and saved the model with the best performance. Subsequently, we evaluated the performance of this best model using the test dataset.

C. Baselines

In this section, we compare the CSCNet with several state-of-the-art methods, including three attention-based methods (DSIFN [16], DTCDSCN [15], and SNUNet [30]) and four transformer-based methods (BIT [22], ChangeFormer [23], CropLand [5], and ICIF [31]). For those methods that expose the training weights, we use the weights provided by the original authors. For those methods that did not disclose the training

TABLE I
COMPARISON RESULTS ON THREE CD TEST SETS

Models	LEVIR-CD Pre. / Rec. / F1 / IoU / OA	WHU-CD Pre. / Rec. / F1 / IoU / OA	GZ-CD Pre. / Rec. / F1 / IoU / OA
DTCDSCN	88.53 / 86.83 / 87.67 / 78.05 / 98.77	91.84 / 89.16 / 90.48 / 82.62 / 99.09	88.19 / 78.38 / 83.00 / 70.93 / 96.92
DSIFN	94.02 / 82.93 / 88.13 / 78.77 / 98.87	96.91 / 73.20 / 83.41 / 71.53 / 98.83	55.70 / 67.41 / 67.00 / 43.99 / 91.74
SNUNet	89.18 / 87.17 / 88.16 / 78.83 / 98.82	91.34 / 85.53 / 88.34 / 79.11 / 98.91	87.55 / 80.05 / 83.64 / 71.87 / 97.00
BIT	89.24 / 89.37 / 89.31 / 80.68 / 98.92	88.71 / 86.27 / 87.47 / 77.73 / 98.81	82.40 / 78.18 / 80.23 / 66.99 / 96.31
MSCANet	89.79 / 87.57 / 88.67 / 79.64 / 98.86	93.16 / 84.50 / 88.62 / 79.56 / 98.95	83.04 / 78.42 / 80.66 / 67.59 / 96.40
ChangeFormer	92.05 / 88.80 / 90.40 / 82.48 / 99.04	88.50 / 85.33 / 86.88 / 76.81 / 98.76	84.59 / 65.23 / 73.66 / 58.30 / 95.53
AMTNet	91.82 / 89.71 / 90.76 / 83.08 / 99.07	91.11 / **89.97** / 90.57 / 82.62 / 99.10	**87.98** / 77.44 / 82.38 / 70.03 / 96.83
ICIF-Net	91.13 / 90.57 / **91.18** / **83.85** / **99.12**	92.83 / 88.70 / **90.77** / **83.09** / **99.13**	89.90 / **80.76** / **85.09** / **74.05** / 97.29
CSCNet_S	**92.54** / 90.68 / 91.61 / 84.52 / 99.15	**94.95** / 90.22 / 93.00 / 86.91 / 99.35	85.54 / 84.78 / 85.16 / 74.17 / **97.17**
CSCNet_L	93.30 / **90.39** / 91.82 / 84.88 / 99.18	96.18 / 90.94 / 93.49 / 87.77 / 99.39	86.02 / 84.66 / 85.34 / 74.37 / 97.39

The highest score is marked in bold and colored in red, the second highest score is colored in blue, and the third highest score is marked in bold. All the scores are described as percentages (%).

TABLE II
PARAMETER SETTINGS FOR CSCNET_S AND CSCNET_L

Models	ResNet	CSCM	TAM	LEVIR-CD		WHU-CD		GZ-CD	
				F1	IoU	F1	IoU	F1	IoU
CSCNet_S	x1 (64, 128, 128) x2 (128, 64, 64) x3 (256, 32, 32)	CSCM_1 (64) CSCM_2 (128) CSCM_2 (256)	TAM_1 (256) TAM_2 (256) TAM_3 (128)	91.61	84.52	93.00	86.91	85.16	74.17
CSCNet_L	x1 (128, 128, 128) x2 (256, 64, 64) x3 (512, 32, 32)	CSCM_1 (128) CSCM_2 (256) CSCM_2 (512)	TAM_1 (256) TAM_2 (256) TAM_3 (128)	91.82	84.88	93.49	87.77	85.24	74.27

All scores are described as percentages (%).

weights, we retrained the model according to the code and guidelines provided by the original author. Especially, for the LVEIR-CD method, we used the model provided by the original author and obtained similar results on our test set as in the original article. However, because the WHU and GZ datasets were cropped by us, and the dataset settings were different for each study group, we had to train these models ourselves to ensure fair comparisons.

1) DTCDSCN [15] is an attention-based method that proposes a twin CNN with a dual-task constraint. The network includes a CD network and two semantic segmentation networks, as well as a dual-attention module.
2) DSIFN [16] is a CD network that uses a multiscale feature concatenation method. This article proposes a deep supervised image fusion CD network that uses a discriminative network with deep supervision to achieve the integrity of the graphic boundaries and density of the internal changes and fuses multilevel deep features with image difference features through an attention mechanism.
3) SNUNet [30] is a multilevel feature fusion method that employs the NestedUNet architecture for CD. The model uses dense connections and deep supervision to improve the recognition ability of intermediate features and enhance the effectiveness of the final features.
4) MSCANet [5] is a feature difference-based method based on the transformer architecture. It proposes a CNN-transformer network with multiscale context aggregation (MSCANet) that combines the advantages of CNN and transformer to achieve efficient and effective CD.
5) BIT [22] is a feature-level difference method based on transformers. It extracts tokens containing rich image features by combining them with semantic tokens and then uses transformers to enhance their contextual modeling capabilities.
6) ChangeFormer [23] is a transformer-based method that uses feature concatenation. The backbone of this method consists entirely of Transformer encoders without the use of CNNs. MLPs are used as decoders and are integrated into the UNet architecture for CD.
7) AMTNet [32] is a multiscale transformer network. AMTNet based on the attention mechanism is proposed, which is especially designed for remote sensing image CD. By utilizing the attention mechanism and the multiscale feature fusion structure, image features at different scales can be effectively captured and utilized to achieve more accurate and robust CD. In particular, the network is able to adaptively process features at multiple scales to gain a deep understanding of changes in remotely sensed images.

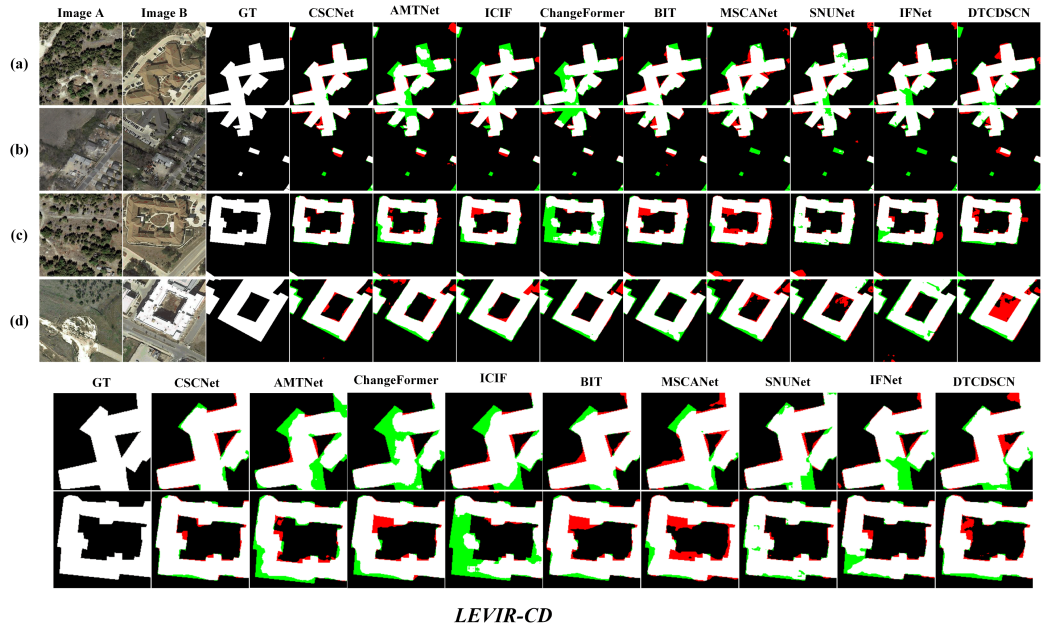

Fig. 4. (a)–(d) Visualization results on the LEVIR-CD dataset. Different colors represent different results, with white representing true positive, black representing true negative, red representing false positive, and green representing false negative. The local detail maps of (a) and (c) are shown at the bottom.

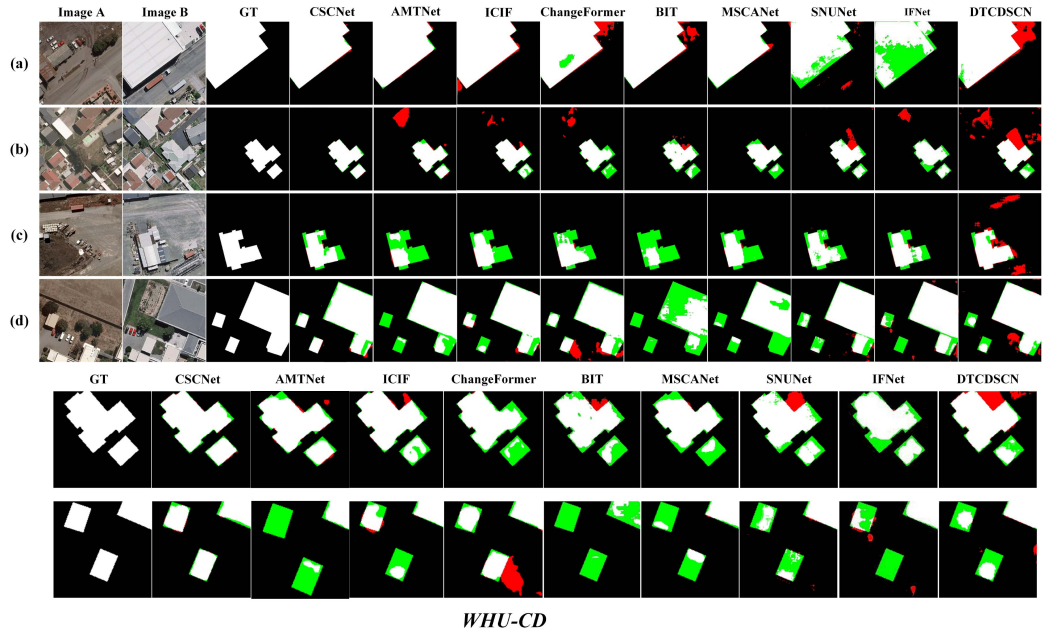

Fig. 5. (a)–(d) Visualization results on the WHU-CD dataset. Different colors represent different results, where white represents true positives, black represents true negatives, red represents false positives, and green represents false negatives. The local detail images of (b) and (d) are shown at the bottom.

8) ICIF [31] proposes a scale-invariant cross interaction and scalewise feature fusion network (ICIF-Net) that combines the advantages of both the CNN and the Transformer for addressing the issue of modeling long-range dependencies in remote sensing CD.

D. Comparison Results

In our comparative experiments, we used the publicly available codes provided by the authors and kept the default and common parameters consistent to ensure fairness. In addition, we set the training epochs to the same value to eliminate the error differences caused by different parameter selections and training rounds, thereby achieving a more accurate comparison of the performance of different methods. Furthermore, this approach helps to avoid unnecessary interference factors during the experiment, ensuring the reliability and reproducibility of the experimental results.

1) Comparison Test Results: In Table I, we present the evaluation results for all the methods on the three datasets, with the

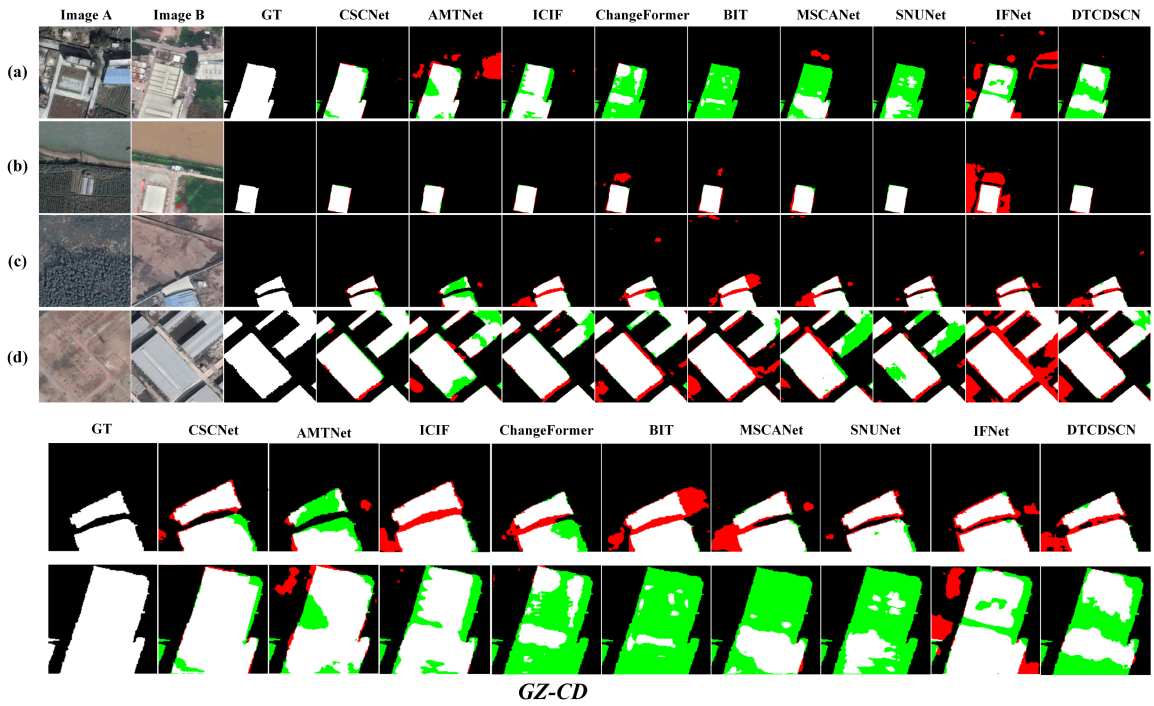

Fig. 6. (a)–(d) Visualization results on the GZ-CD dataset. Different colors represent different results, where white represents true positives, black represents true negatives, red represents false positives, and green represents false negatives. The local detail images of (a) and (d) are shown at the bottom.

highest values indicated in bold. In the comparative experiments, we showcase two testing models, namely, CSC_S and CSC_L, and their internal parameter settings are listed in Table II. By comparing the results in the table, it can be observed that our CSCNet outperforms significantly in most of the metrics. While we may not achieve the highest precision, the comprehensive criterion for evaluating methods is typically the F1-score, which combines information from both precision and recall. In terms of the F1-score metric, the CSCNet improves by 0.43%, 2.23%, and 0.07% compared to the second-best method on the three datasets, respectively. In addition, CSC_L demonstrates improvements of 0.21%, 0.49%, and 0.18% over CSC_S. These results indicate that our deep model exhibits significant performance advantages in remote sensing CD tasks, particularly in key metrics such as F1-score and IoU. These metrics are crucial for remote sensing image CD, as they provide a comprehensive assessment of the model's performance, balancing precision and recall. Therefore, our approach holds great potential for applications in remote sensing image CD.

In the GZ-CD dataset, we noticed that the performance of CSCNet was not as outstanding as on the other two datasets. This could be attributed to differences in the labeling approach used for the GZ-CD dataset compared with the other two datasets. In the LEVIR-CD and WHU-CD datasets, labels were created for each building as well as edge details, whereas in the GZ-CD dataset, the labels treated the entire building cluster area as a single target region and did not individually label the edges of each building within the cluster. This difference in labeling methodology may have resulted in the CSCNet not performing as well in capturing details on the GZ-CD dataset compared with the other datasets.

TABLE III
PARAMETER AND FLOP RESULTS FOR ALL THE METHODS ON THE THREE DATASETS AND THE F1-SCORE AND IOU VALUES ON EACH DATASET

Models	Params(M)	FLOPs(G)	LEVIR-CD		WHU-CD		GZ-CD	
			F1	IoU	F1	IoU	F1	IoU
DTCDSCN	41.07	13.21	87.67	78.05	85.03	73.96	83.00	70.93
DSIFN	50.71	82.35	88.13	78.77	74.23	59.03	67.00	43.99
SNUNet	12.03	54.88	88.13	78.83	85.26	74.31	84.25	72.79
BIT	3.55	10.59	89.31	80.68	84.90	73.75	80.23	66.99
MSCANet	16.42	14.77	88.67	79.64	79.64	66.17	80.66	67.59
ChangeFormer	41.03	202.87	90.40	82.48	81.82	69.24	73.66	58.30
ICIF-Net	25.83	25.27	91.18	83.52	90.77	83.09	85.09	74.05
CSCNet_S	18.11	19.00	91.61	84.52	93.00	86.91	85.16	74.17
CSCNet_L	56.57	25.09	91.82	84.88	93.49	87.77	85.34	74.27

2) Visualization Results: In this section, we visualize the results on the LEVIR-CD (see Fig. 4), WHU-CD (see Fig. 5), and GZ-CD (see Fig. 6) datasets, using different colors to represent true positives (TP—white), true negatives (TN—black), false positives (FP—red), and false negatives (FN—green). This visualization allows for an intuitive comparison of the differences between the CSCNet and other advanced methods. In LEVIR-CD and WHU-CD, it is evident that our CSCNet performs better in capturing edge details. Particularly, in WHU-CD, our method excels in recognizing the completeness of targets, and it also outperforms other methods in recognizing multiple targets. In GZ-CD, our method exhibits fewer noise artifacts, as shown in (c) and (d) in the images in Figs. 4–6. We also perform well in building recognition, with fewer green areas indicating false negatives compared to other methods. These visual results further highlight the advantages of CSCNet in handling high-resolution remote sensing image CD tasks.

TABLE IV
IMPACT OF DIFFERENT MODULE COMBINATIONS ON MODEL PERFORMANCE ON THREE DIFFERENT BUILDING CD DATASETS (LEVIR-CD, WHU-CD, AND GZ-CD)

Module	DO-Conv	CSCM	TAM	LEVIR-CD F1	LEVIR-CD IoU	WHU-CD F1	WHU-CD IoU	GZ-CD F1	GZ-CD IoU
CSCNet_S	✓	×	×	90.81	83.47	90.77	84.21	81.17	69.43
CSCNet_S	×	×	✓	90.53	82.97	90.26	83.74	80.44	68.82
CSCNet_S	×	✓	×	91.07	83.60	91.53	84.67	82.11	70.23
CSCNet_S	✓	×	✓	91.30	84.02	92.04	85.25	84.10	72.56
CSCNet_S	×	✓	✓	91.42	84.19	92.75	86.49	84.45	73.08
CSCNet_S	✓	✓	×	91.48	84.30	92.51	86.06	83.26	71.31
CSCNet_S	✓	✓	✓	91.61	84.52	93.00	86.91	85.16	74.17

Modules include DO-Conv, CSCM, and TAM. We use "✓" to indicate that the module is included and "×" to indicate that the module is not included. The performance of each module combination is measured by F1-score and IoU. We sort in ascending order by the F1 index on the dataset LEVIR-CD.

3) Efficiency and Performance of the Model: To compare the efficiency of various methods, we list the parameter count, floating point operations (FLOPs), and test results on the datasets for all the compared methods in Table III. In terms of the parameter count, our CSC_S model has a lower parameter count compared to recent Transformer-based methods and does not significantly increase computational overhead. However, our method achieves significant performance improvements on various datasets. In the CSC_L model, we increase the parameter count, but there is no significant increase in computational cost. Despite the larger parameter count, the FLOP performance remains good, and the increase in parameter count does not lead to a substantial increase in computational overhead. In addition, we use DO-Conv to improve ResNet, significantly reducing computational complexity. The differences in ResNet will be compared in detail in Section IV-E. These results indicate that our method achieves significant performance improvements while maintaining computational efficiency across various datasets.

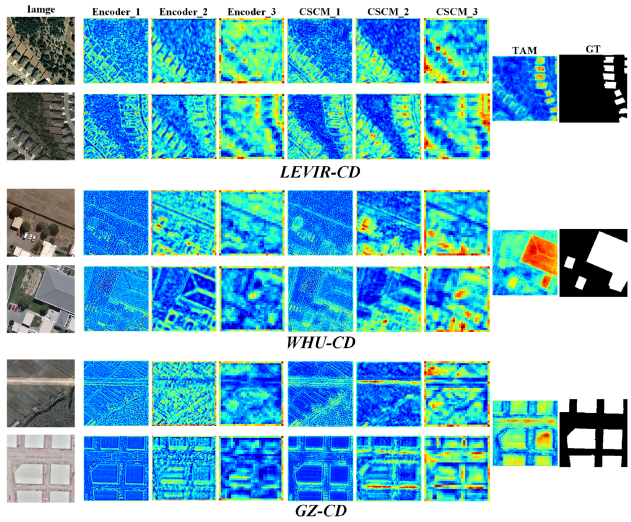

Fig. 7. Visualization results of the various stages of the CSCNet. It includes the feature difference maps for each stage of the encoder, each stage of the CSCM, and the TAM output.

E. Ablation Experiments

In the ablation studies, we dissected three pivotal modules of the CSC-S model, including the encoder module, the CSCM, and the TAM. Initially, in the encoder ablation experiment, we contrasted the original ResNet with the enhanced ResNet equipped with DO-Conv. Substantial performance enhancements were realized across all three datasets. The experimental outcomes are presented in Table IV.

In the ablation trials of the CSCM, its impact on model performance was assessed by partial removal of the CSCM. The findings illustrated that, across all three datasets, the absence of the CSCM led to a notable performance decline. This further accentuates the vital role of the CSCM in capturing multiscale features and orchestrating feature relationships.

In the ablation tests of the TAM, we substituted the TAM with convolution blocks, which resulted in a performance dip, particularly pronounced in the GZ-CD dataset. This signifies the substantial contribution of the TAM in leveraging the Transformer for contextual modeling and its importance for model performance.

Moreover, simultaneous ablation experiments were conducted on both the CSCM and the TAM. The results demonstrated a notable performance decline across all three datasets when these two critical modules were removed or replaced. This further underscores the critical role of the CSCM and the TAM in the contextual modeling of the CSC method.

Overall, the results from the ablation experiments indicate that both the CSCM and the TAM play crucial roles in the CSCNet method. They contribute to enhancing the model's feature representation and contextual modeling capabilities, thereby ameliorating the performance of remote sensing image CD.

In Table IV, we observed that the inclusion of TAM seemed to lead to a performance drop in the experiments on the WHU-CD dataset. Regarding this, the TAM aims to aggregate features by leveraging the Transformer structure to capture long-term dependencies and contextual information. However, on the WHU-CD dataset, possibly due to the relatively simple spatial relationships between targets or less pronounced long-term dependencies, the TAM might not have provided much aid and may even have introduced additional noise, thus affecting performance. We plan to further investigate and optimize the TAM in future work to ensure that it provides stable performance improvements across different datasets and tasks.

F. Network Visualization

For a better visual interpretation of the crucial stages in our CSCNet, we conducted visualizations of the intermediate layers of our network, as displayed in Fig. 7. This figure illustrates the visualization results of the CSCNet at various stages. It encompasses the different phases of the encoder, each stage of the CSCM, and the feature difference maps produced by the TAM. From these visuals, it is evident that the model's attention varies across certain key feature regions, which are highly relevant to the task at hand. Specifically, within the CSCM and the TAM, we can intuitively observe how the model processes and focuses on vital contextual information at different stages, thereby achieving a richer and more precise feature representation. These visualization outcomes not only offer us an in-depth understanding of the working mechanism of CSCNet but also attest to its efficiency and robustness in handling complex tasks.

V. Conclusion

In this article, we introduced a deep learning model called CSCNet. Our model improved upon the traditional ResNet by employing the DO-Conv technique to reduce computational complexity while enhancing model expressiveness. We also introduced the CSCM, allowing features at various stages of the encoder to interact locally and globally, adaptively focusing on target features at different scales. In addition, we designed the TAM as the decoder to better aggregate multiscale features and establish stronger contextual connections. On three different building CD datasets (LEVIR-CD, WHU-CD, and GZ-CD), our CSCNet demonstrated significant advantages, achieving F1-score evaluation metrics of 91.61%, 93.00%, and 85.16%, respectively. This highlights CSCNet's outstanding performance and effectiveness in handling high-resolution remote sensing image CD tasks. Despite some advancements our model made in multiscale feature fusion and contextual modeling, there are still some limitations. First, our model has not been rigorously compared with already published methods that address multilevel feature fusion, and its effectiveness may still need further validation. Second, although our model has achieved some success in reducing computational complexity, it may still face challenges when dealing with large-scale or more complex datasets. Moreover, our model mainly focuses on building CD tasks and has not yet been validated on other remote sensing image CD tasks. Compared to already published methods that address multilevel feature fusion, our model attempts to realize adaptive interaction and contextual association between features through the CSCM and TAM components. This to some extent solves the problems of fixed feature fusion strategies that traditional methods may encounter. However, further research is needed in the future to better understand and optimize multiscale feature interaction and contextual modeling strategies to further improve the performance and effectiveness of remote sensing image CD. Our research results offer a promising approach in the field of remote sensing image CD and provide valuable insights for further research applying deep learning to this domain. By incorporating key components such as DO-Conv, CSCM, and TAM, our model addresses challenges related to multiscale and contextual modeling, providing a powerful tool for high-quality building CD.

References

[1] S. Iino, R. Ito, K. Doi, T. Imaizumi, and S. Hikosaka, "Generating high-accuracy urban distribution map for short-term change monitoring based on convolutional neural network by utilizing SAR imagery," *Proc. SPIE*, vol. 10428, 2017, Art. no. 1042803.

[2] B. Peng, Z. Meng, Q. Huang, and C. Wang, "Patch similarity convolutional neural network for urban flood extent mapping using bi-temporal satellite multispectral imagery," *Remote Sens.*, vol. 11, no. 21, 2019, Art. no. 2492.

[3] P. P. De Bem, O. A. de Carvalho Junior, R. F. Guimarães, and R. A. T. Gomes, "Change detection of deforestation in the Brazilian Amazon using landsat data and convolutional neural networks," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 901.

[4] C. Mucher, K. Steinnocher, F. Kressler, and C. Heunks, "Land cover characterization and change detection for environmental monitoring of pan-Europe," *Int. J. Remote Sens.*, vol. 21, nos. 6/7, pp. 1159–1181, 2000.

[5] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multi-scale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[6] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[7] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[8] S. Jiao, X. Li, and X. Lu, "An improved Ostu method for image segmentation," in *Proc. 8th Int. Conf. Signal Process.*, 2006, pp. 16–19.

[9] G. Byrne, P. Crapper, and K. Mayo, "Monitoring land-cover change by principal component analysis of multitemporal landsat data," *Remote Sens. Environ.*, vol. 10, no. 3, pp. 175–184, 1980.

[10] J. Chen, P. Gong, C. He, R. Pu, and P. Shi, "Land-use/land-cover change detection using improved change-vector analysis," *Photogrammetric Eng. Remote Sens.*, vol. 69, no. 4, pp. 369–379, 2003.

[11] 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.

[12] J. Chen et al., "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2020.

[13] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[14] 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.

[15] 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.

[16] 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.
[17] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.
[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] M. Gong, Y. Yang, T. Zhan, X. Niu, and S. Li, "A generative discriminatory classified network for change detection in multispectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 321–333, Jan. 2019.
[20] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.
[21] S. Saha, F. Bovolo, and L. Bruzzone, "Change detection in image time-series using unsupervised LSTM," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020.
[22] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 8005205.
[23] 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, doi: 10.1109/IGARSS46834.2022.9883686.
[24] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.
[25] X. Song, Z. Hua, and J. Li, "Remote sensing image change detection transformer network based on dual-feature mixed attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416.
[26] J. Cao et al., "DO-Conv: Depthwise over-parameterized convolutional layer," *IEEE Trans. Image Process.*, vol. 31, pp. 3726–3736, 2022.
[27] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[28] K. Song, F. Cui, and J. Jiang, "An efficient lightweight neural network for remote sensing image change detection," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 5152.
[29] S. Li and L. Huo, "Remote sensing image change detection based on fully convolutional network with pyramid attention," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4352–4355.
[30] 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, 2021, Art. no. 8007805.
[31] 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.
[32] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 202, pp. 599–609, 2023.
[33] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[34] 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.
[35] 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.

Yiyang Zhao received the bachelor's degree in computer science and technology from the School of Computer Science and Technology, Yantai Institute of Technology, Yantai, China, in 2019. He is currently working toward the master's degree in computer science and technology with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai.

His research interests include computer graphics, computer vision, and image processing.

Xinyang Song received the bachelor's degree in computer science and technology from the School of Computer Science and Technology, Yantai Institute of Technology, Yantai, China, in 2019. He is currently working toward the master's degree in computer science and technology with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai.

His research interests include computer graphics, computer vision, and image processing.

Jinjiang Li received the B.S. and M.S. degrees from the Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shandong University, Jinan, China, in 2010, all in computer science.

From 2004 to 2006, he was an Assistant Research Fellow with the Institute of Computer Science and Technology, Peking University, Beijing, China. From 2012 to 2014, he was a Postdoctoral Fellow with Tsinghua University, Beijing. He is currently a Professor with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China. His research interests include image processing, computer graphics, computer vision, and machine learning.

Yepeng Liu received the bachelor's and Ph.D. degrees in computer science and technology from the School of Computer Science and Technology, Shandong University, Jinan, China, in 2014 and 2020, respectively.

He is currently a Teacher with Shandong Technology and Business University, Yantai, China. His research interests include image superresolution, image denoising, and medical image processing.