Enhanced Self-Attention Network for Remote Sensing Building Change Detection

Shike Liang, Zhen Hua, and Jinjiang Li

***Abstract*—The self-attention mechanism can break the limitation of the receptive field, model in a global scope, and extract global information efficiently. In this work, we propose a lightweight remote sensing building change detection model (ESACD). In the encoder, we use the enhanced self-attention layer, CoT layer, instead of the normal convolution operation. The CoT layer fuses the dynamic context with the static context. Compared with the ordinary convolutional layer, this strategy can fully mine the local features between the input keys to dynamically enhance the feature representation. Subsequently, we use dual attention to further mine the low-frequency information and high-frequency information of the images and the semantic features of interest to the model. Dual attention consists of the HiLo attention mechanism and the Tokenizer attention mechanism. HiLo extracts high-frequency information and low-frequency information through two branches. In the high-frequency branch, nonoverlapping windows are applied to the features for self-attention. In the low-frequency branch, average pooling is first performed on features before self-attention. After Tokenizer attention extracts the feature tokens that the model is interested in, it encodes its information and, then, converts the tokens into pixel-level features. Tokenizer attention realizes the efficient extraction of features and enhances the representation ability of the model. Finally, we fuse feature information to enhance the fluidity of information and improve accuracy. Through our experiments on two change detection datasets, ESACD has better performance than other state-of-the-art methods while maintaining fewer parameters.**

***Index Terms*—Change detection (CD), remote sensing building images, self-attention.**

I. Introduction

THE task of remote sensing change detection has developed rapidly recently and has attracted the attention of a large number of researchers. Given a bitemporal image of the same area, the purpose of building change detection is to detect the area where the building changes in the images. And by using a binary image to represent the change map, that is, the area with architectural changes is represented by white, and the area without architectural change detection is represented by black. In particular, note that there are correlated and uncorrelated changes in bitemporal images. Relevant changes refer to changes in the building, and nonrelevant changes refer to changes due to sunlight [1], soil [2], and seasonal changes. The model must have the ability to distinguish relevant changes from nonrelated changes to avoid misjudgment of nonrelated changes as architectural changes, resulting in a decrease in model accuracy. The development of high-resolution remote sensing architectural images has played a vital role in the application of change detection. Now remote sensing image change detection has been widely used in various fields such as urban sprawl detection [3], [4], [5], postdisaster reconstruction [1], [6], and land use detection [7]. In this work, we propose a lightweight change detection model for remote sensing building images.

Many fields of computer vision are currently affected by deep learning, and research progress has been greatly accelerated [8]. In the deep-learning model, the essence of CNN is to use the convolution kernel training parameters to learn the local feature information of pixels. By continuously stacking and expanding the receptive field of the convolutional layer, the convolutional layer can learn more semantic information [9], [10]. Due to the existence of CNNs local mechanism, CNN has the characteristics of local spatial locality and translation invariance. The popularity of transformers can be seen from the original ViT [11] and numerous ViT variants, such as Swin TransFormer [12] and Fastformer [13]. ViT divides the image into multiple patches and uses the linear embedding sequence of these patches as the input of transformers. In transformers, patches are first expanded into 2-D sequence blocks and positional embeddings are added to encode positional information. Then, the overall information is sent to the transformer encoder, and the self-attention operation is used to perform remote semantic modeling to learn global feature information. The reason why transformers can show superior performance mainly depends on the self-attention mechanism. The self-attention mechanism is a variant of the attention mechanism, which reduces the dependence on external information and has better performance in capturing the internal correlation of data or features. We believe that the way CNN and self-attention mechanisms extract features are complementary forms, and we have been trying to fuse the characteristics of the two. CNN uses the convolution kernel to mine the local pixel set centered on a certain pixel, which can efficiently extract the information in the local semantic image. However, in the process of

Manuscript received 4 March 2023; revised 23 April 2023; accepted 17 May 2023. Date of publication 25 May 2023; date of current version 7 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62272281, Grant 62002200, Grant 62202268, and Grant 61972235, in part by the Shandong Natural Science Foundation of China under Grant ZR2021MF107 and Grant ZR2022MA076, and in part by the Yantai Science and Technology Innovation Development Plan under Grant 2022JCYJ031. *(Corresponding author: Zhen Hua.)*

Shike Liang and Jinjiang Li are with the School of Information and Electronic Engineering, Institute of Network Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: 2021420057@sdtbu.edu.cn; lijinjiang@gmail.com).

Zhen Hua is with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: huazhen@sdtbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3278726

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

learning features, with the accumulation of convolutional layers, it is easy to fail to fit and lack the ability to learn global features. Transformer-based remote sensing change detection methods are good at establishing long-distance dependencies. Compared with expanding the receptive field in CNN, transformers can perform remote modeling more directly and efficiently. However, the transformer-based remote sensing change detection method faces the problems of many parameters and a large amount of calculation and cannot efficiently process large-scale image features. Therefore, many researchers are now paying more attention to the lightweight direction of transformers.

What we need to know is that the image contains rich frequency information, and the low frequency and high frequency play different roles. The authors in [14] and [15] analyzed the role of frequency in computer vision tasks. Low frequencies in an image usually represent global information, such as global structure, color, etc.; high frequencies in an image represent detailed features, such as sharp edges. Based on this high- and low-frequency information, researchers have proposed numerous solutions for image rescaling [16], generalization [17], image superresolution [18], [19], and neural network compression [20], [21]. At present, there is not much work to apply the extraction of different frequencies of images to remote sensing change detection tasks. We hope to make full use of the information carried by different frequencies in the images to enhance the model's ability to represent features and simulate global relationships. Two attention heads are adopted in [22] to model local and global features. Among them, the first attention head is used to capture the local features of the image and focus on the local area in the image by calculating the attention weight on the low-resolution feature map of the image. The second attention head is used for global feature modeling, by calculating the attention weight on the high-resolution feature map of the image, so as to focus on the global information of the entire image. Affected by the excellent performance in [22], we introduce the HiLo attention mechanism to mine global and fine features by using low-frequency and high-frequency information in the images. For the high-frequency and low-frequency information in the images, different branches are used to encode the multihead self-attention mechanism. In the high-frequency branch, the high-resolution feature maps are extracted through the local window self-attention mechanism of a certain number of attention heads for high-frequency feature extraction. In the low-frequency branch, the global self-attention mechanism of the remaining attention heads is used to encode low-frequency information on the downsampled feature maps, which can reduce the computational load of the model and maintain a faster speed. The low-frequency branch is then fused with the high-frequency branch to simulate global and local features. HiLo is more friendly to hardware requirements. Experiments have proved that HiLo has achieved excellent performance and is more memory efficient.

In this work, we use the enhanced self-attention CoT layer [23] to replace the ordinary convolutional layer and use dual attention to enhance the expressiveness of the model. Li et al. [23] propose to use a context-aware transformer network to improve visual recognition performance, which can capture contextual information in images and use this information to improve the effect of the model. CoT first utilizes 3×3 convolutions to encode the static context and then utilizes two consecutive 1×1 convolutions to dynamically learn the multihead attention matrix. Finally, the learned weight matrix is multiplied by the input value to obtain a dynamic context representation, which is fused with the static context to obtain the final output. We use a typical U-Net architecture to extract semantic features of different resolutions in bitemporal images. In the encoder in U-Net, we use the CoT layer to extract the local feature information in the images; in the decoder, we subtract the two-channel features to learn the feature difference and then use the CoT layer to decode the features. Compared with ordinary convolution, CoT is more able to strengthen the connection of input keys to context. Then, we utilize the dual-attention mechanism including HiLo attention and Tokenizer attention [24] to enhance the modeling ability of the model. Chen et al. [24] proposed to extract only the features of interest in the image, convert pixel-level information into a small number of tags, and provide a better input representation for subsequent deep-learning models. The dual-attention mechanism model can enhance the learning of context and capture the connection between key pixels. Finally, the learned features are fully fused and predicted to get the best results.

In summary, our contributions are threefold.

1) In this work, we propose a lightweight change detection model for remote sensing building images. We use the CoT layer instead of ordinary convolution operation to mine dynamic context and static context in parallel, so that dynamic context can alleviate the limitation of receptive field on model effect in convolution operation with the assistance of static context. At the same time, the fusion of dynamic context and static context also enables the model to learn richer contextual semantic information and enhances the feature expression ability of the model. After our extensive experiments on two datasets, our model has excellent performance with a small number of parameters.

2) In order to further mine the global structure information and fine feature information carried by the low frequency and high frequency in the images, we introduce the HiLo attention mechanism. By giving the low-frequency branch and the high-frequency branch a certain number of attention heads to extract the information carried by the two, the ability of the model to simulate the feature relationship is enhanced. In the high-frequency branch, nonoverlapping windows are used for self-attention mechanism operation, and the average pooling operation is performed on the feature maps in the low-frequency branch to reduce image resolution. Through the above two strategies, the HiLo mechanism has low complexity and high throughput.

3) We propose a lightweight ESACD implementation for change detection. It follows the U-Net architecture, uses a parameter-shared encoder to process bitemporal images, and then passes the bitemporal feature differences to the decoder for decoding. We also convert changes of interest to the model into compact tokens using Tokenizer attention. Subsequently, use Tokens encoder and Tokens decoder for context modeling and project back to pixel-level features to supplement details.

II. Related Work

A. Traditional CD Method

In the remote sensing change detection network based on the traditional method, it mainly includes the following three steps: 1) data preprocessing, 2) feature extraction, 3) discrimination and classification. Data preprocessing is to reduce spurious changes due to data reasons. Commonly used data preprocessing techniques mainly include georeferencing and radiometric correction, both of which are mainly used to solve the problem of the geographical location of remote sensing images and the radiation differences generated when different sensors acquire remote sensing images. In the following process of feature extraction and discriminative classification, the model converts the information in the images into other forms and then analyzes the differences to obtain the final difference map. Common traditional methods mainly include clustering and PCA methods [25], [26]. Due to the development of technology, it is also easier to obtain high-quality remote sensing images. High-quality remote sensing images contain a large number of texture features and spatial structure information. The use of traditional remote sensing change detection methods has been unable to meet the needs of people to achieve high-precision prediction results. With the hot momentum of deep learning, many researchers pay more attention to deep learning, expecting to use deep learning to achieve better change detection results.

B. Deep Learning CD Method

Feature extraction is divided into pixel-based [27], [28], object-based [29], [30], [31] and feature-based [32], [33] based on the basic unit of processing. Pixel-level-based methods are similar to traditional image-transformation-based methods [34]. A neural network is used to convert the images into a deep feature space and then perform a pixel-by-pixel comparison of the deep features to distinguish changing pixels. For example, Zhang et al. [35] learned deep features and then maps bitemporal features to a 2-D polar domain for differential change recognition. Saha et al. [36] used a pretrained neural network to extract semantic features change vector analysis (CVA) and compare features to generate a change map. Object-level-based methods mainly use superpixels to segment building objects to accurately identify building boundaries. The features are then learned using a deep CNN for change recognition. For example, Zhang et al. [37] extracted fine features and superpixel branches through parallel neural networks to strengthen the retention of boundary information of building objects. Finally, the detailed features are added to the superpixel branch feature maps, and the feature details of different resolutions are fused to achieve the final change prediction. Feature-based methods tend to utilize end-to-end neural networks to learn features in a supervised manner for change detection.

CD method based on pure CNN: At present, the mainstream structure based on the neural network is to use the Siamese structure. According to whether parameter sharing is used, it can be divided into pure Siamese network and pseudo-Siamese network. The pure Siamese network is able to measure the similarity of bitemporal images by extracting common features of them. Undoubtedly, this is more suitable for the change detection task. The pseudo-Siamese network extracts its semantic features through two branches. This increases the flexibility of the model, but at the same time inevitably increases the number of trainable parameters and complexity. Later, on the basis of the Siamese network, Daudt et al. [38] used the fully connected layer FCN architecture for the first time to realize the remote sensing change detection task and proposed three benchmark models: 1) FC-EF, 2) FC-Siam-Diff, and 3) FC-Siam-Conc. This makes an important contribution to the research of change detection in the future. FC-EF first added the bitemporal images directly and then implemented the remote sensing change detection network by applying the skip-connected U-Net architecture. FC-Siam-Diff learned bitemporal image similarity by applying a pure Siamese network with a parameter-sharing encoder. FC-Siam-Conc achieved change detection by subtracting the bitemporal features shared by parameters to obtain the difference between them. It should be noted that the strategy of directly adding bitemporal images in FC-EF lacks the advantage of comparing similar features over the dual-channel Siamese network, so few researchers use this strategy. Then, in order to further improve the accuracy of predicting the boundaries of building regions, Liu et al. [39] proposed a deep pure Siamese dual-task building change detection network. The network used semantic segmentation tasks to assist the model to further improve the accuracy of building recognition before predicting differences but this also imposed requirements on the dataset. At the same time, the network also introduced an attention module to strengthen the learning channel and spatial features and improved the loss function. Zhang et al. [35] proposed a deeply supervised image fusion network, IFNet. First, the pseudo-Siamese encoder is used to learn features of different resolutions, and then, deep supervision is used to distinguish feature differences. Finally, spatial attention and channel attention are used to fully integrate different features. However, due to the existence of the pseudo-Siamese network, the IFNet model had a large amount of parameters.

Transformer-based CD method: After transformers have achieved excellent results in the NLP field. Many researchers have begun to try to apply transformers to the field of computer vision [40]. ViT [11] was proposed in 2020 and showed the surprising performance of transformers. However, ViT also has obvious shortcomings. The premise of ViT showing good results is that there are enough data for pretraining; otherwise, transformers will not be able to break through the limitation of lack of inductive bias. Subsequently, many variants of ViT began to emerge, such as Swin Transformer [12], Pyramid Vision Transformer [41], Lite Transformer [42], and DelighT [43]. Transformer models proved its effect in the CV field. In the field of change detection, many researchers have introduced transformers and achieved good results. Chen et al. [24] used the ResNet dual channel with parameter sharing for feature extraction and then sent the features to the encoder by block to simulate the global relationship. Then, the decoder is used to complete the feature relationship for prediction. In [44], Bandara and Patel proposed a transformer-based Siamese network, which

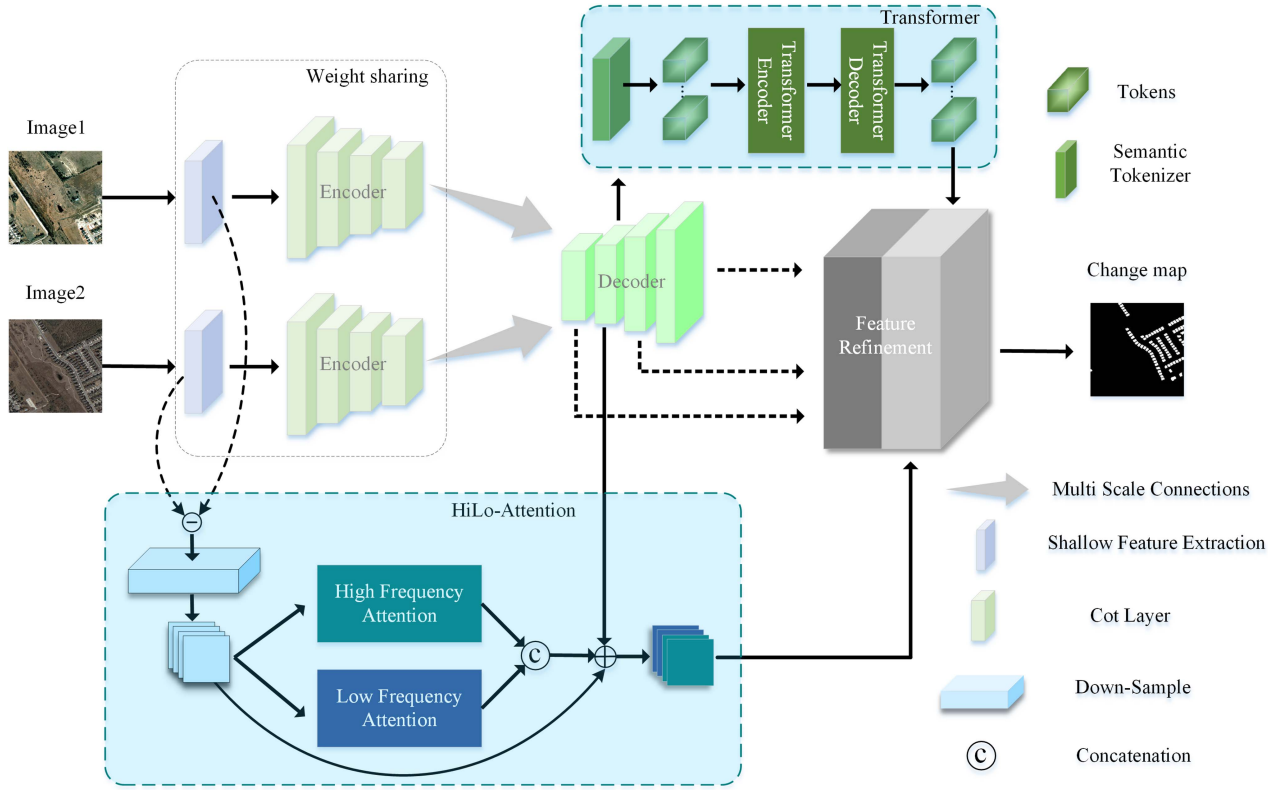


Fig. 1. ESACD of the overall network architecture.

mainly included a transformer encoder to learn dual-channel image features and then used a feature difference module for each size to compare the differences between the two. Finally, a lightweight MLP is used to decode and predict the final change result. ChangeFormer showed excellent results. Liu et al. [45] proposed an end-to-end transformer-based network, which combined prior extraction and contextual fusion together by learning prior-aware transformers. Wang et al. [46] proposed a network for scene change detection. The work in [46] is mainly composed of CNN backbone, Siamese ViT, and prediction head. CNN backbone is mainly used to extract feature information. Siamese ViT is used to establish the global semantic relationship and the long-term context of the model, making the model more robust to noise changes. The prediction head is mainly composed of transposed convolutions to restore the original scale feature relationship and then predict the change result.

Attention-based CD method: At the same time, it is also common to use attention-based models in the field of remote sensing [47]. In the field of remote sensing change detection, Chen et al. [48] proposed a new Siamese neural network to solve the CD algorithm problem by combining the spatial attention mechanism with the location attention mechanism. Zheng et al. [49] utilized high-frequency attention HFAB-guided symmetric networks. HFAB aims to enhance the model's ability to acquire high-frequency information of buildings, and it mainly consists of two stages, namely spatial attention mechanism and high-frequency enhanced attention. The spatial attention mechanism first guides the model to focus on the building area of interest, and then high-frequency enhanced attention is used to highlight the high-frequency information in the input feature map, which can better detect the edges of changing buildings. Chen et al. [50] proposed a fully convolutional Siamese remote sensing change detection neural network based on dual attention composed of spatial attention and channel attention. Long-range dependencies are captured by a dual-attention mechanism for more discriminative feature representations.

In this work, we propose a lightweight and efficient neural network for remote sensing change detection. The backbone of the network is a U-Net architecture based on shared parameters. It should be noted that we use a CoT layer that combines self-attention and convolution in the U-Net architecture to alleviate the limitations of the receptive field of the convolutional layer. And this will fully mine the global and local information to improve the expressive ability of the model. We then use a dual-attention module to improve the modeling ability of the model.

III. Method

In this work, we propose a change detection network ESACD for remote sensing, and the network structure diagram is shown in Fig. 1. Our model is mainly composed of the classic U-Net architecture based on self-attention mechanism, dual attention, and feature fusion module (FFM). The input of the model is a bitemporal image $I^1, I^2 \in R^{H\times W\times 3}$. Next, we elaborate on each part of the model in detail.

The detailed logic details of our ESACD model are presented in Algorithm 1.

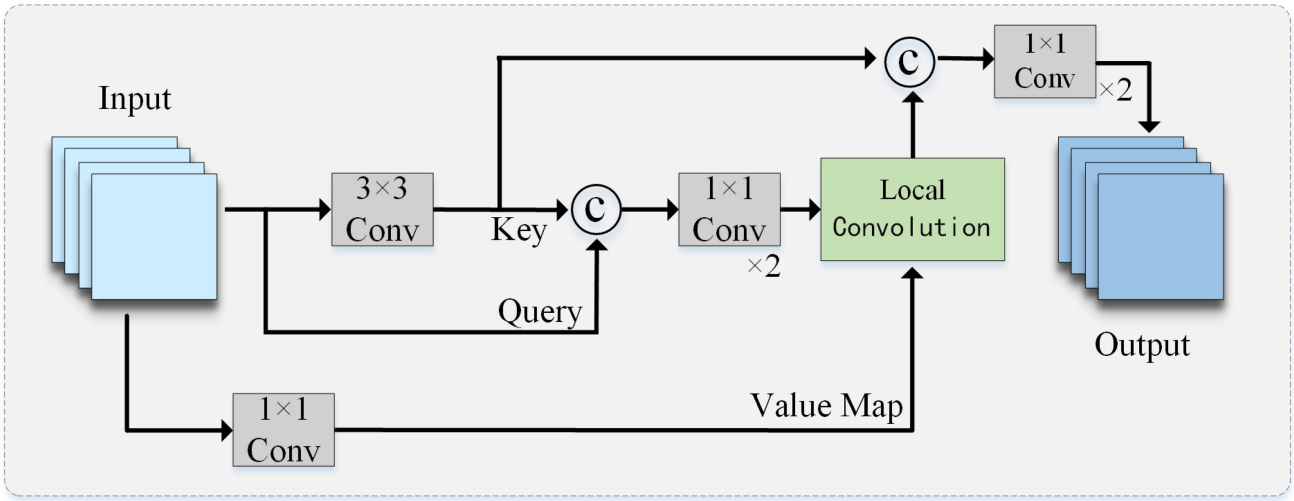


Fig. 2. CoT layer module structure.

Algorithm 1: Inference of ESACD-Based Model for Change Detection.

Input: I^1, I^2
Output: change map
step1: Shallow Feature Extraction
$F^{\text{sf1}}, F^{\text{sf2}} = \text{Shallow Feature Extraction}(I^1, I^2)$
step2: Use the CoT layer in the encoder to first encode the bitemporal features
$A^{\text{out}} \in \{A^{\text{out1}}, A^{\text{out2}}, A^{\text{out3}}, A^{\text{out4}}\} = \text{CoT}(F^{\text{sf1}})$
$P^{\text{out}} \in \{P^{\text{out1}}, P^{\text{out2}}, P^{\text{out3}}, P^{\text{out4}}\} = \text{CoT}(F^{\text{sf2}})$
step3: In the decoder, the dimensions of the dual-temporal features are subtracted and decoded
$M^{\text{out}} \in \{M^{\text{out1}}, M^{\text{out2}}, M^{\text{out3}}, M^{\text{out4}}\} = \text{CoT}(A^{\text{out}} - P^{\text{out}})$
step4: Subtract the extracted shallow features to perform high- and low-frequency attention operations and then add them to M^{out2}
$M^{\text{out}}_{\text{HiLo}} = \text{HiLo}(F^{\text{sf1}} - F^{\text{sf2}})$
step5: Enhance model modeling ability with the Tokenizer attention mechanism
$M^{\text{out}}_{\text{Token}} = \text{Tokenizer}(M^{\text{out2}})$
step6: Fusion and prediction of feature
change map $= \text{Feature Fusion}(M^{\text{out}}, M^{\text{out}}_{\text{HiLo}}, M^{\text{out}}_{\text{Token}})$
return change map

A. CoT Layer

In the classic U-Net architecture, given the input bitemporal images I^1, I^2, we first use shallow feature extraction to increase the feature dimension to learn more feature information and get feature maps $(R^{H\times W\times 3})$. Then, we pass the feature maps into the U-Net network composed of encoder and decoder shared by parameters to extract semantic features of different resolutions. Different from the previous classic U-Net, we use the self-attention mechanism of the CoT layer to replace the conventional convolutional layer in the U-Net architecture to further learn contextual key features. This can avoid introducing additional branches to further learn the feature context and cause additional parameters. The CoT layer structure is shown in Fig. 2. In the face of input feature maps, the CoT layer will extract static and dynamic context features in parallel. The CoT layer learns static context keys by encoding input keys using a 3×3 convolutional layer. The static context guarantees the local features learned by taking advantage of the spatial locality of the convolution and variability such as translation. In the dynamic context branch, the relationship between each key and query is used to fuse them to obtain the attention weight matrix between key and query. The attention weight matrix is then multiplied with value to achieve a dynamic contextual representation of the input. This process fully integrates the information provided by key, query, and value to realize the dynamic mining of the global context. Finally, the dynamic context and the static context are fused to realize self-attention learning. This design pays more attention to the rich context between adjacent keys than the traditional self-attention mechanism. At the same time, this also enables the dynamic context to promote the learning of visual representation under the guidance of the static context and effectively alleviates the limitation effect of the convolutional layer receptive field on learning global features. This process can be expressed as

$$\text{Static} = \text{Conv3} \times 3\,(\text{Input}) \tag{1a}$$

$$QK = Cat\,([\text{Static}, \text{Input}]) \tag{1b}$$

$$V = \text{Conv1} \times 1\,(\text{Input}) \tag{1c}$$

$$\text{Dynamic} = \text{LocalConvolution}\,[\text{Conv1} \times 1(QK), V] \tag{1d}$$

$$\text{Output} = \text{Conv1} \times 1\,(\text{Cat}[\text{Static}, \text{Dynamic}]) \tag{1e}$$

where $\text{Conv3} \times 3$ means the convolution kernel is 3 and $\text{Conv1} \times 1$ means the convolution kernel is 1.

B. Dual Attention

In this work, we use Tokenizer attention [24] and HiLo attention [22] in parallel to perform long-range modeling and

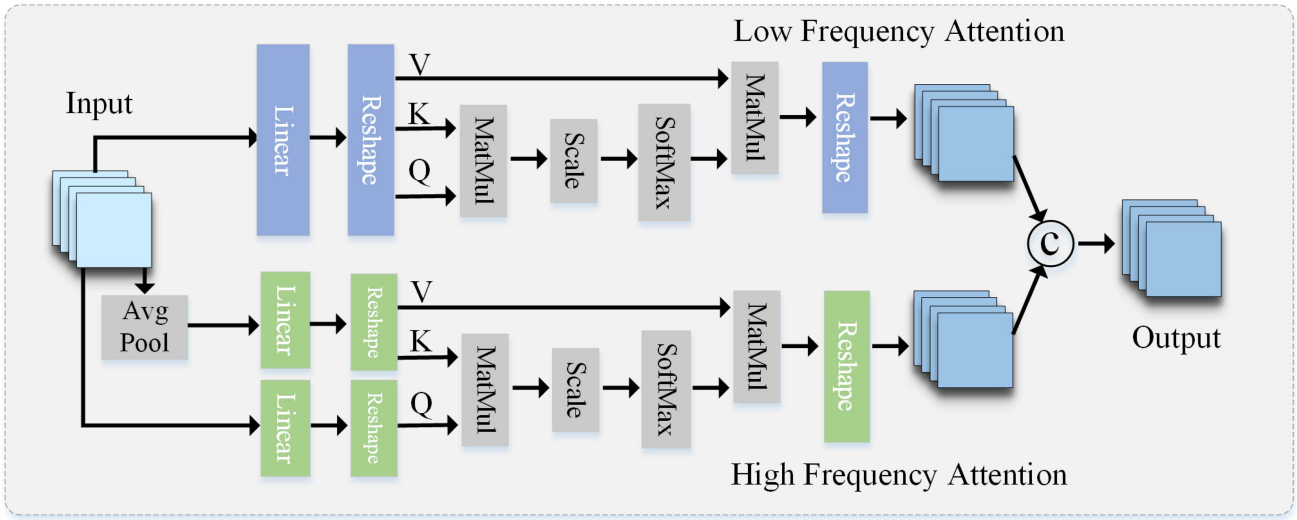


Fig. 3. High- and low-frequency attention structure.

extract high-frequency and low-frequency information to enhance feature representation. Our dual-attention module consists of Tokenizer attention and HiLo attention. We use dual attention to enhance the model's ability to express features and further mine the high- and low-frequency features in the image to carry and encode information. Tokenizer learns the feature representation of the model's interest in the image and enhances the feature expression ability of the model. In our work, Tokenizer learns multiple sets of Tokens with varying features by utilizing spatial attention, so we call this module Tokenizer attention. Tokenizer can effectively improve the feature representation ability of the model while maintaining a small amount of calculation. HiLo attention uses two branches to learn the high-frequency and low-frequency features contained in the image in different forms. Dual attention plays the role of a strengthening link in ESACD, and its two ends are, respectively, connected to the U-Net architecture composed of CoT layer and FFM. Our experiments have proved that dual attention can guarantee a certain strengthening effect by using Tokenizer attention and HiLo attention. The more intuitive effect can be observed in Fig. 10.

1) HiLo Attention: In images, high frequencies capture local fine details while low frequencies focus more on global features. Therefore, we introduce HiLo attention to encode high-frequency features and low-frequency features through two branches and then fuse them to improve the efficiency of feature extraction. As shown in Fig. 3, in the face of feature maps $F^{\mathrm{sf}} \in R^{256\times256\times64}$ two branches are used to unlock the high frequency and low frequency of the attention layer, respectively. In the high-frequency branch, the high frequency is encoded through the self-attention in each local window; in the low-frequency branch, the global relationship is modeled by performing self-attention on each window's average pooled low-frequency key and each query of the input feature. In order to achieve better efficiency, different numbers of heads are set for the high-frequency and low-frequency branches. Keep the same number N of heads in MSA and assign $(1-\alpha)N$ for high-frequency feature encoding, and the other αN for low-frequency feature extraction. Through such a strategy, HiLo can maintain low complexity and high throughput, which helps to reduce model parameters.

In detail, in the high-frequency branch, learning detail features through local window self-attention is more efficient than standard MSA, so as to achieve high-frequency attention features. In the low-frequency branch, the low-frequency signal is obtained by pooling the window, and then, the remaining heads are used to model the relationship between the low-frequency signal key and the query in the feature map. Since the pooling operation reduces the complexity of query and key, the complexity of low-frequency branches is also reduced. Finally, the results obtained by the two branches are fused using the concatenation operation to realize the interaction between low-frequency information and high-frequency information. Benefiting from the absence of time-consuming operations such as window sliding and recursion in the two branches, HiLo runs very fast on the device. Expressed as

$$X_{\mathrm{init}} = \mathrm{DownSample}\left(M^{\mathrm{out2}}\right) \tag{2a}$$

$$\mathrm{High}^{\mathrm{out}} = \mathrm{High}\left(X_{\mathrm{init}}\right) \tag{2b}$$

$$\mathrm{Low}^{\mathrm{out}} = \mathrm{Low}\left(X_{\mathrm{init}}\right) \tag{2c}$$

$$M^{\mathrm{out}}_{\mathrm{HiLo}} = \mathrm{Cat}\left(\mathrm{High}^{\mathrm{out}}, \mathrm{Low}^{\mathrm{out}}\right) + X_{\mathrm{init}} \tag{2d}$$

where High$(\cdot)$ represents the high-frequency feature extraction branch, and Low$(\cdot)$ represents the low-frequency feature extraction branch. The DownSample module uses a convolutional layer with a convolution kernel of 3×3. After the HiLo attention mechanism, we can get the output result $M^{\mathrm{out}}_{\mathrm{HiLo}} \in R^{64\times64\times64}$.

2) Tokenizer Attention: We introduce Tokenizer to learn the feature representation of the model's interest in the image and enhance the feature expression ability of the model. In our work, Tokenizer learns multiple sets of Tokens with varying features by utilizing spatial attention, so we call this module

Tokenizer Attention. Tokenizer can effectively improve the feature representation ability of the model while maintaining a small amount of calculation. During the positive transmission of Tokenizer attention, a token extractor is first used to convert feature maps into semantic tokens to represent the changed features of interest to the model in bitemporal images. Similar to tokens in natural language processing, the token extractor divides the entire feature map into several tokens, and each token corresponds to a label vector. In this process, facing feature maps $F_1 \in R^{64\times64\times64}$. First, a convolutional layer is used to obtain L semantic groups, and then, the spatial attention calculation and SoftMax operation are performed on each semantic group to obtain the spatial attention map. Finally, the attention map is used to calculate the weighted average sum of the pixels in F_1 to obtain semantic tokens $T_1 \in R^{L\times C}$. This process can be expressed as

$$T_1 = \left(\text{SoftMax}\left(\text{Conv1}\times 1\left(F_1; W\right)\right)\right)^T F_1 \tag{3}$$

where $W \in R^{L\times C}$ represents a learnable kernel.

The Tokenizer encoder is composed of N layers of multihead self-attention (MSA) and multilayer perceptron (MLP) blocks. And the prenorm residual unit (PreNorm) is used for normalization before each layer of MSA and MLP. The Tokenizer encoder is to model the context of tokens and simulate the global relationship. At layer k, MSA accepts tokens from layer $k-1$ and obtains Q, K, and V for multihead self-attention operations. The specific formula is as follows:

$$Q = T^{k-1}W^q \tag{4a}$$

$$K = T^{k-1}W^k \tag{4b}$$

$$V = T^{k-1}W^v \tag{4c}$$

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4d}$$

where $W^q, W^k,$ and W^v are the learnable parameters of three linear projection layers.

MSA processes multiple attention heads in parallel and uses the concatenation operation to fuse multiple attention heads to obtain the final value. MSA plays an important role in the Tokenizer encoder capturing the long-term relationship of tokens. We then need to project the token representations back to pixel space to obtain pixel-level features. To this end, we use Tokens decoder to refine pixel features. Tokens decoder consists of M layers of multihead cross attention (MA) and MLP blocks. Except for using MA, the rest of the structure is similar to Tokens encoder. We use MA to avoid the model to heavily compute dense relationships between F_1 pixels. It should be noted that in MSA, the query, key, and value are derived from the same input sequence, whereas in MA, the query is from the image features, and the key and value are from the tokens. The detailed process can be expressed by the formula as follows:

$$\text{token} = \text{Token\_Extractor}\left(M^{\text{out2}}\right) \tag{5a}$$

$$M^{\text{mid}}_{\text{Token}} = \text{Token\_Encoder}(\text{token}) \tag{5b}$$

$$M^{\text{out}}_{\text{Token}} = \text{Token\_Decoder}(M^{\text{mid}}_{\text{Token}}). \tag{5c}$$

After the above process, we can get the output result $M^{\text{out}}_{\text{tokens}} \in R^{64\times64\times64}$.

TABLE I
MORE DETAILS ON THE OVERALL STRUCTURE OF THE ESACD MODEL

Stage	ESACD	Output
SFE	[Conv 3,1]×3	256×256
Encoder 1	[CoT 32,3]×2	256×256
Encoder 2	[Max pool 2,2] [CoT 32,64,3] [CoT 64,64,3]	128×128 128×128 128×128
Encoder 3	[Max pool 2,2] [CoT 64,128,3] [CoT 128,128,3]	64×64 64×64 64×64
Encoder 4	[Max pool 2,2] [CoT 128,256,3] [CoT 256,256,3]	32×32 32×32 32×32
Decoder	[ConvTranspose2d 2,2] [CoT 3]	[32,64,128,256]
HiLo	[High and Low frequency attention]	64×64
Tokenizer	[Transformer encoder transformer decoder]	64×64
FFM	[Upsample] [Conv 3,1]×3	256×256

C. Feature Fusion Module

We use FFM to fully fuse features of different resolutions. The FFM structure is shown in Fig. 4. FFM first uses the Upsample module to unify the size to 256 × 256. Before performing this process, first, add the feature maps that have undergone HiLo attention and Tokenizer attention. For the Upsample module, we use the linear interpolation algorithm bilinear to achieve upsampling. Next, we apply a convolution layer fusion feature with a 3 × 3 convolution kernel to the feature maps after the concatenation operation to obtain the final prediction result change map $R^{256\times256\times2}$. This process can be used with the formula expressed as

$$W^{\text{mid}} = \text{Upsample}\left(M^{\text{out}}, M^{\text{out}}_{\text{HiLo}} + M^{\text{out}}_{\text{Token}}\right) \tag{6a}$$

$$W^{\text{out}} = \text{Cat}\left(\text{Conv3}\times 3\left(W^{\text{mid}}\right)\right) \tag{6b}$$

$$\text{change map} = \text{Pred}\left(\text{Conv3}\times 3\left(\text{Conv3}\times 3\left(W^{\text{out}}\right)\right)\right) \tag{6c}$$

where Pred represents prediction head, and the number of change map channels is 2.

In Table I, we present more details on the overall structure of the ESACD model in a tabular form. It should be noted that the parameters following Conv, Max pool, and ConvTranspose2d are the convolution kernel size and stride, and the parameters following CoT are the input dimension, output dimension, and convolution kernel size.

D. Loss Function

Cross-entropy (CE) loss is used to evaluate the difference between bitemporal features. CE is a widely used loss function for remote sensing change detection. The model first calculates the error between the predicted value and the original label and then uses the loss function to backpropagate the error, calculates the gradient of the error relative to the network parameters, moves the parameters in the opposite direction of the gradient, and continuously updates the parameters to reach the lowest

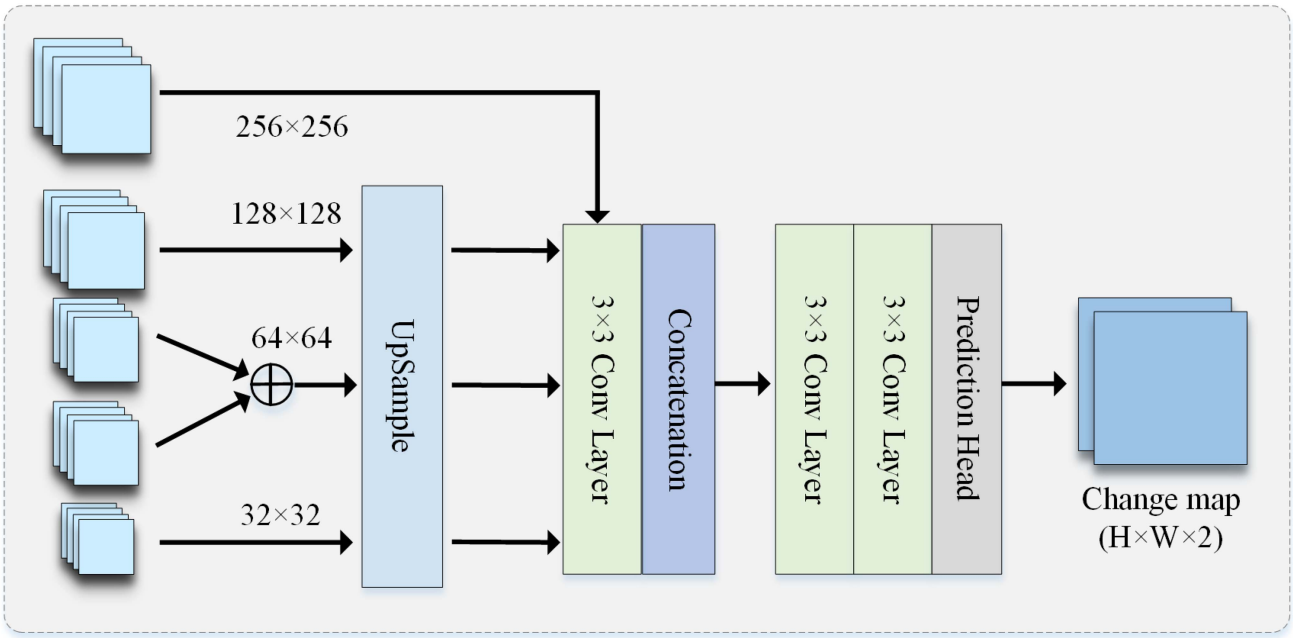


Fig. 4. FFM structure.

value of the loss. The formula of the loss function is as follows:

$$L = \frac{1}{H_0 \times W_0} \sum_{h=1,w=1}^{H,W} l\left(P_{hw}, Y_{hw}\right) \tag{7}$$

where Y_{hw} is the label for the pixel at location (h, w), $l(P_{hw}, y) = -log(P_{hwy})$ is the CE loss and H_0 and W_0 represent the length and width of the original image, respectively.

IV. Experiment

A. Experiment Details

Our proposed model is implemented through Pytorch. In the training process, we apply normal data augmentation to the input image patches, including random flipping, random rescaling, random color dithering, random cropping, and Gaussian blur. Our ESACD is trained in an end-to-end fashion using AdamW as the optimizer with a weight decay of 0.01 and a beta value of (0.9, 0.999) to optimize the model. We set the batch size as B = 256. The loss used for training is CE, the initial learning rate is set to 0.0001 and it gradually decays to 0. We train it until the model fits perfectly.

B. Datasets and Metrics

In this experiment, we use the WHU-CD [51] dataset and the Google Data [52] dataset for comparative experiments and ablation experiments.

WHU-CD [51] is a public building CD dataset. It includes two aerial images with 0.075-m spatial resolution, both of which were taken in Christchurch, but at different times, one in 2012 and the other in 2016. The original image size in WHU-CD is 32507 × 15354 pixels. We will crop it to a nonoverlapping 256 × 256 image block. Also, we randomly divide it into training set, validation set, and test set in the ratio of 8:1:1.

The Google Data [52] collect 19 pairs of VHR images of seasonal changes with a resolution of 0.55 m. The images document changes in the suburbs of Guangzhou, China, between 2006 and 2019. Image sizes range from 1006 × 1168 to 4936 × 5224 pixels. We will crop it to a nonoverlapping 256 × 256 image block. We randomly divide it into the training set and the test set in the ratio of 10:1.

In this experiment, in order to compare the effect of our model with other models in many aspects, we use F1, IoU, over accuracy (OA), precision, and recall as evaluation indicators. Among them, since F1 and IoU are commonly used as evaluation indicators in image segmentation tasks, we choose F1 and IoU as our main evaluation indicators, and the rest of the indicators are used as auxiliary indicators. F1 is used to measure the accuracy of the binary classification model. IoU represents the correlation between the predicted value and the real value, the higher the correlation, the higher the value. Precision represents the proportion of samples identified as positive that are correctly identified. Recall represents the proportion of all positive samples that were correctly identified in the prediction

$$F1 = 2 \times \frac{\text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \tag{8a}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{8b}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{8c}$$

$$\text{Rec} = 2 \times \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8d}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8e}$$

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

C. Comparison With Recent Methods

In order to verify the effectiveness of our model, we choose methods that have performed well in recent years to compare with them. These methods are mainly divided into three categories: The first category is CNN-based methods, including FC-EF [38], FC-Siam-Di [38], and FC-Siam-Conc [38]. These methods all adopt U-Net architecture to extract semantic information at different levels. FC-EF, FC-Siam-Di, and FC-Siam-Conc are classic models in the field of remote sensing building change detection, which have great significance for the future development of this field. The difference among the three lies in the processing method of bitemporal images. FC-EF added the input bitemporal images and then transfers them to the model; FC-Siam-Di subtracted bitemporal features to extract features; FC-Siam-Conc used parameter sharing to directly perform concatenation operations on bitemporal features. The second category is attention-based methods, including DSIFN [35]. DSIFN proposed a multidimensional image fusion network that fused bitemporal features in a disparity discrimination network to ensure effective interaction of bitemporal features through an attention mechanism. In addition, depth supervision is used to effectively improve the ability to discriminate changing pixels. The third category is transformer-based methods, including BIT, ChangeFormer, Transcd, and PaFormer. These methods used transformer's self-attention mechanism for remote semantic modeling and have achieved excellent results. Chen et al. [24] proposed a transformer-based approach for CD. It slices the input image into multiple tokens and models the context based on the tokens. In addition, BIT has good performance in terms of efficiency and accuracy without using complex structures. The work in [46] is a transformer-based neural network for scene CD. Wang et al. [46] improve the recognition of regions of interest in images and improve robustness to noise changes by establishing long-term contextual relationships. The work in [45] is an end-to-end transformer-based network, which combines prior extraction and contextual fusion together by learning prior-aware transformer. Bandara and Patel [44] used the hierarchical transformer encoder as well as the lightweight MLP decoder to add sensory fields to enhance context shaping and feature representation capabilities.

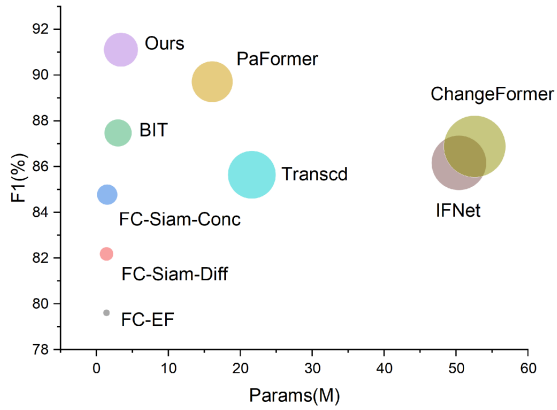
The comparison results of our model with other excellent models on the dataset WHU-CD and Google-CD test set are shown in Tables II and III. In the WHU-CD dataset, our main evaluation metrics F1 and IoU outperform the second place by 1.6% and 2.9%, respectively. In the Google-CD dataset, our main evaluation metrics F1 and IoU outperform the second place by 0.7% and 1.2%, respectively. It is clear that the performance of our proposed ESACD outperforms the rest of the models on both datasets. To further demonstrate that our ESACD is lightweight, we show in Fig. 5(a) and (b) how ESACD compares with other methods on the WHU-CD dataset on the metrics F1 and number

TABLE II
COMPARISON OF RESULTS ON THE WHU-CD DATASET

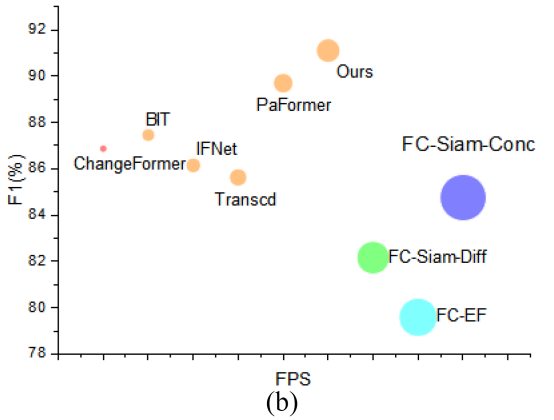
Network	Precision	Recall	F1	IoU	OA
FC-EF[38]	83.07	76.41	79.60	66.11	98.11
FC-Siam-diff[38]	83.64	80.77	82.18	69.74	98.31
FC-Siam-conc[38]	87.98	81.78	84.77	73.56	98.58
IFNet[34]	88.10	84.29	86.16	75.68	98.69
BIT[24]	88.71	86.27	87.47	77.73	98.81
ChangeFormer[44]	88.50	85.33	86.88	76.81	98.76
TransCD[46]	82.26	89.30	85.64	74.88	98.56
PaFormer[45]	85.99	93.77	89.71	81.34	98.96
ESACD	90.72	91.50	91.11	83.67	99.14

TABLE III
COMPARISON OF RESULTS ON THE GOOGLE-CD DATASET

Network	Precision	Recall	F1	IoU	OA
FC-EF[38]	82.67	58.96	68.83	52.47	94.89
FC-Siam-diff[38]	74.52	56.62	64.34	47.43	93.99
FC-Siam-conc[38]	76.94	69.49	73.03	57.51	95.08
IFNet[34]	70.32	56.87	62.88	45.83	93.57
BIT[24]	87.98	74.99	80.97	68.02	96.62
ChangeFormer[44]	76.20	70.43	73.20	57.73	95.06
TransCD[46]	85.89	75.42	80.32	67.11	96.46
PaFormer[46]	87.37	78.18	82.52	70.24	96.83
ESACD	84.74	81.49	83.08	71.06	96.82



(a)



(b)

Fig. 5. (a) Comparison of parameter quantities between ESACD and other CD methods on the WHU-CD dataset. (b) Comparison of FPS between ESACD and other CD methods on the WHU-CD dataset.

TABLE IV
PARAMETER DETAILS OF ESACD AND OTHER COMPARISON METHODS

Model	Param(M)	Flops(G)
FC-EF	1.4	3.5
FC-Siam-Diff	1.4	4.7
FC-Siam-Conc	1.5	5.3
IFNet	50.4	27.6
BIT	3.0	8.8
ChangeFormer	52.6	20.9
Transcd	21.6	28.3
PaFormer	16.1	10.9
ESACD	3.5	18.4

of parameters and FPS, respectively. We can see that ESACD has excellent performance and the parameter quantity is relatively small. Although the number of ESACD parameters is not the least, its evaluation indicators are very outstanding. In Fig. 5(b), we can observe that ESACD has a higher FPS while maintaining the highest F1. It should be noted that the higher the value of FPS, the faster the running speed of the model. Meanwhile, we show the specific parameter data of ESACD and other comparative methods in Table IV.

Figs. 6 and 7 show the renderings of our model and other excellent models on the dataset WHU-CD and Google-CD test sets, respectively. White represents TP, black represents TN, green represents FP, and pink represents FN.

In group a of Fig. 6, we can observe that ESACD has the fewest misjudgment pixels, which are very close to Label. In group b, Transcd and ESACD can more accurately identify architectural change objects but ESACD has fewer misjudged pixel regions than Transcd. In group c, FC-EF, FC-Siam-Di, and FC-Siam-Conc have a serious problem of blurred boundary information. ESACD can accurately judge the characteristics of building boundaries and more accurately identify building areas. In group d, it can be clearly found that ESACD has a higher accuracy rate.

In group a of Fig. 7, we can find that other models cannot accurately identify the subject of architectural change while ESACD can identify objects relatively accurately but there is a problem of fuzzy boundary information. In group b, BIT and ESACD perform well, and other models have too many misjudged pixel areas. In group c, ESACD has the least FN pixel area and can identify the changing subject more completely. In group d, ESACD shows better results than other models.

Overall, our ESACD performs well, with the ability to recognize large building changes, and can accurately identify semantic information of building boundaries. I think that dual attention strengthens the global modeling ability of our model, and the model can learn more similar architectural features and enhance the feature representation ability. In Figs. 8 and 9, we show the local detail images with good performance in Figs. 6 and 7 to feel the effect of ESACD more intuitively.

D. Ablation Experiment

In this section, we will verify various parts of our model to observe the impact of this module on our model. First, we will build four different variants based on whether to add Tokenizer Attention and HiLo Attention, ESACD (base), ESACD (TA), ESACD (HiLo), and ESACD (TA+HiLo). We use these four variants. Experiments are carried out on the WHU-CD dataset, and the main evaluation indicators F1 and IoU are used to evaluate the experimental effect. The experimental results are shown in Table V.

TABLE V
ABLATION EXPERIMENT OF MODULES IN ESACD

Network	Param	F1	IoU	F1	IoU
ESACD(base)	3.08	89.26	80.60	79.52	66.00
ESACD(TA)	3.42	90.36	82.41	82.22	69.81
ESACD(HiLo)	3.15	90.24	82.26	81.95	69.41
ESACD(TA+HiLo)	3.49	91.11	83.67	83.08	71.06

TABLE VI
ABLATION EXPERIMENT OF COT LAYER MODULE IN ESACD

Model	Precision	Recall	F1	IoU	OA
Convolution	89.11	91.24	90.16	82.08	99.04
CoT Layer	92.48	88.82	90.61	82.84	99.11

From Table V, we can observe that Tokenizer attention and HiLo attention play a positive role in the training of our model. At the same time, they spent as little param overhead as possible while improving the model effect, especially HiLo attention.

CoT Layer: In order to verify the effectiveness of the CoT layer, we use ordinary convolution to replace the CoT layer to observe the comparative effect of the two. The specific results are shown in Table VI. We can see that the effect of the model using the CoT layer is significantly higher than that of using ordinary convolution. We think this shows that the strategy of the CoT layer learning dynamic and static context is effective.

Dual Attention: In order to verify the effectiveness of our dual-attention module, we chose the attention module used in STANet [53] and DASNet [50] networks to replace our dual-attention module on the WHU-CD dataset. The details of these two models are as follows.

STANet proposed a remote sensing transformation detection method based on spatial and temporal attention mechanisms. First, the features in the bitemporal image are extracted through feature extraction, and then, the context information and sequence information are extracted by using the spatial and temporal attention mechanisms, and the self-attention calculation is performed on these information, and finally, the spatial and temporal feature information is weighted and fused and predicted to obtain final result.

DASNet is a twin network composed of two fully convolutional neural networks and utilizes spatial attention and channel attention mechanisms to better focus on changing information. Among them, the spatial attention mechanism can adaptively adjust the contribution of different regions of the image to the model; the channel attention module can assign weights to different channels according to the importance of feature maps.

The experimental results are shown in Table VII. From Table VII, we can see that after replacing our dual attention with the attention mechanism in STANet and DASNet, the effect of the model is obviously reduced. The window size of the dual-attention module in STANet is fixed, so there may be certain errors in the recognition of changes in longer time

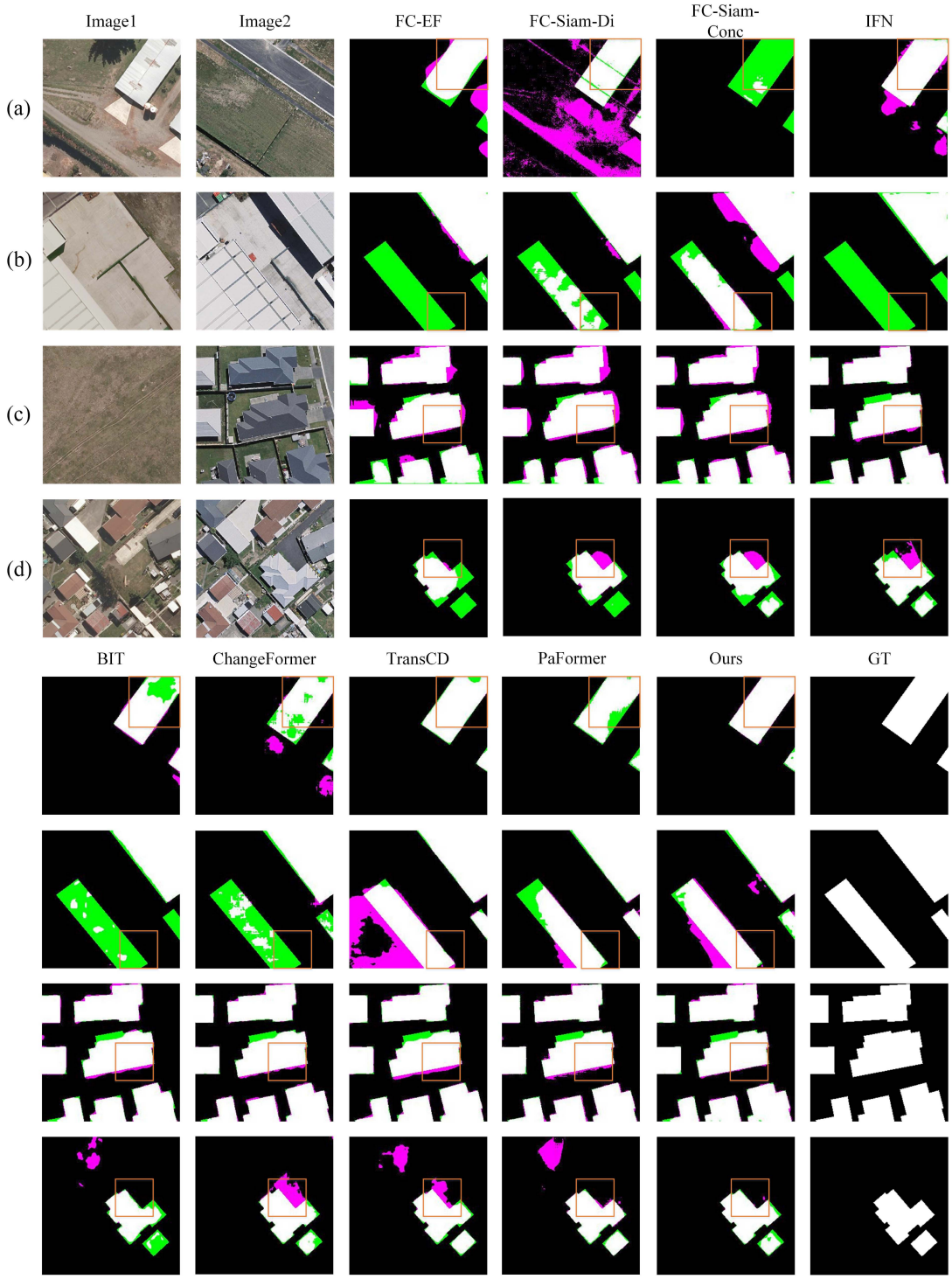


Fig. 6. Qualitative comparison on the WHU-CD dataset. (a), (b), (c), and (d) respectively represent four different sets of bi-temporal images selected from the WHU-CD test set to compare our method with other excellent methods.

scales. This can lead to poor performance in change detection scenarios. We think that the dual-attention module proposed in DASNet imitates the style of self-attention to learn all token similarity features. This will inevitably introduce some noise, which will affect the training effect. And our Tokenizer attention can extract the most interesting target of the model for feature representation, improving the learning efficiency of the model. This further illustrates the effectiveness of our dual attention.

E. Parameter Analysis

In order to further enhance the effect of the model, we conduct parameter analysis on the value of the attention head α in HiLo attention, the layers N and M of Tokens encoder and Tokens decoder, aiming to select the most suitable value for our model to enhance the feature expression of the model ability. First, we select 0.1, 0.3, 0.5, 0.7, and 0.9 for α, respectively, train

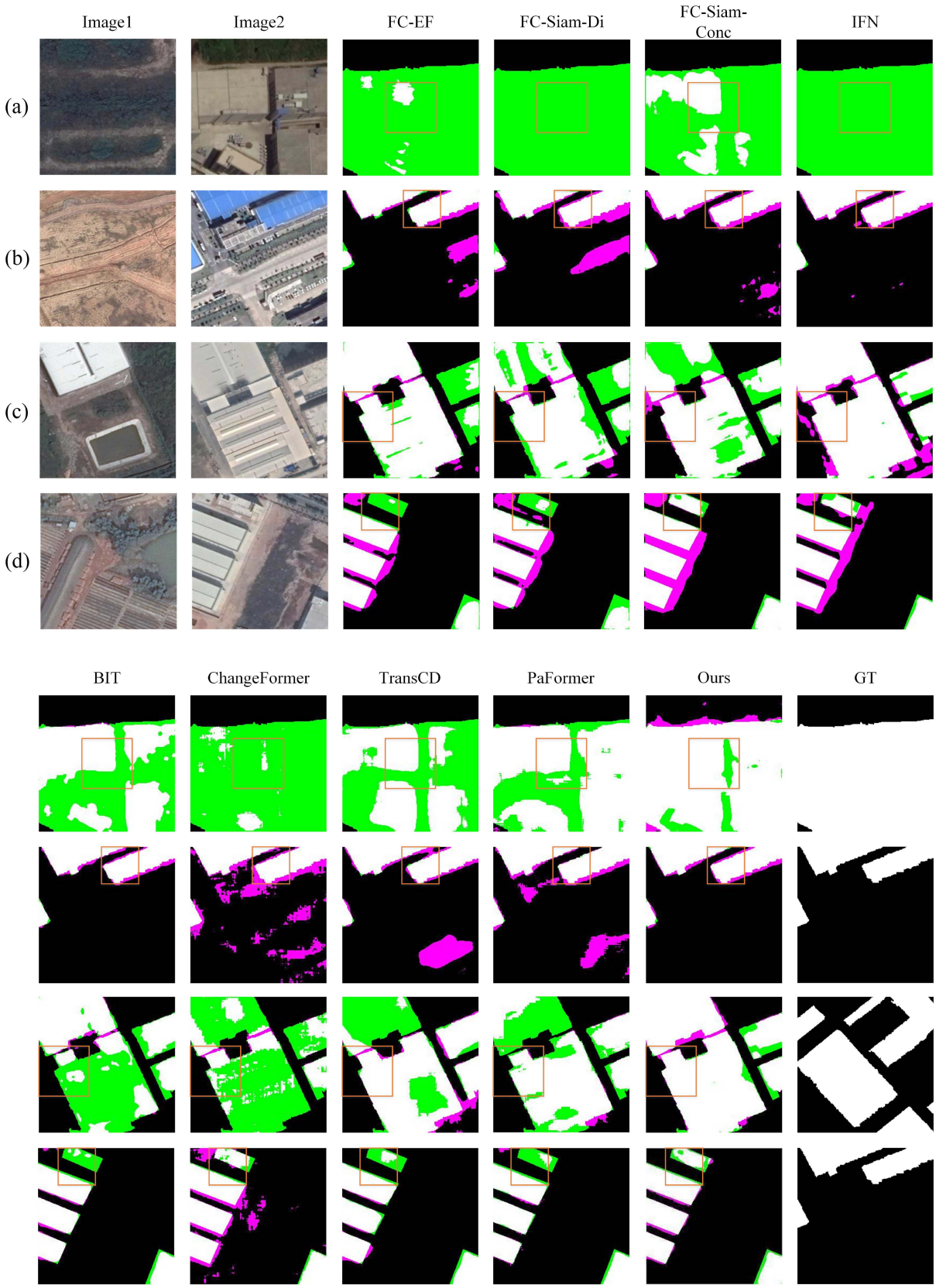


Fig. 7. Qualitative comparison on the Google-CD dataset. (a), (b), (c), and (d) respectively represent four different sets of bi-temporal images selected from the Google-CD test set to compare our method with other excellent methods.

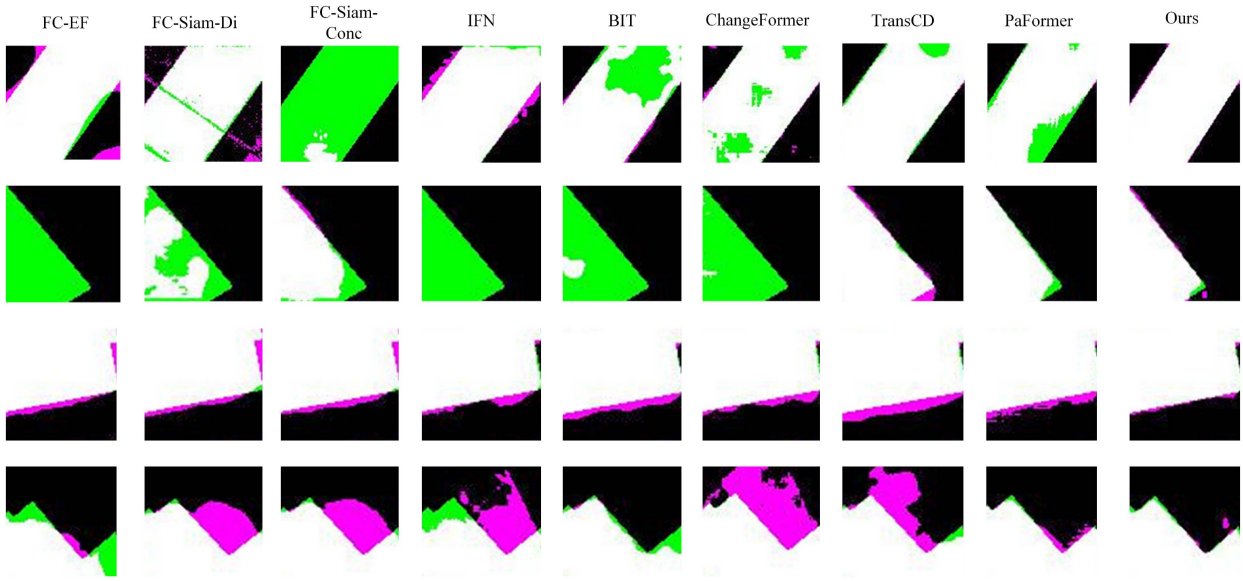


Fig. 8. Detailed demonstration of the WHU-CD test results.

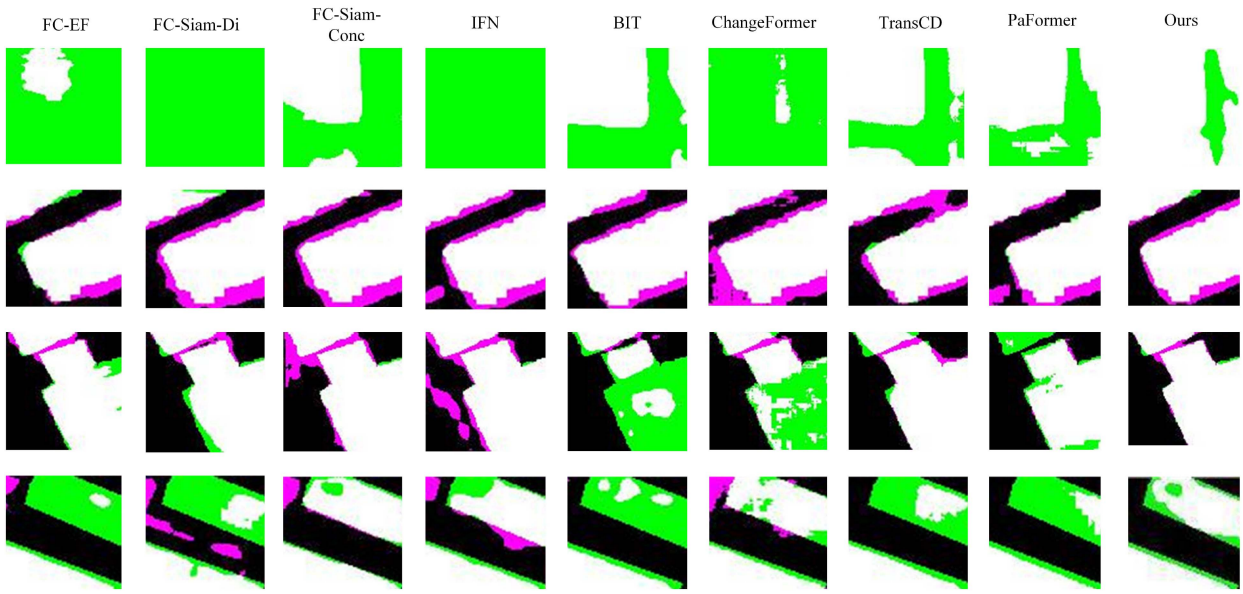


Fig. 9. Detailed demonstration of the Google-CD test results.

TABLE VII
ABLATION EXPERIMENT OF DUAL-ATTENTION MODULE IN ESACD

Model	Precision	Recall	F1	IoU	OA
DASNet	91.67	89.87	90.76	83.08	99.12
STANet	88.14	89.65	88.89	80.00	98.92
ESACD	90.72	91.50	91.11	83.67	99.14

TABLE VIII
ANALYSIS EXPERIMENT OF $\aleph$ PARAMETER IN ESACD ON MODEL EFFECT ON THE WHU-CD DATASET

α	Precision	Recall	F1	IoU	OA
0.1	90.02	91.25	90.63	82.87	99.09
0.3	88.51	92.19	90.31	82.33	99.05
0.5	92.65	89.04	90.81	83.17	99.13
0.7	90.04	91.03	90.53	82.71	99.08
0.9	87.87	90.66	89.24	80.58	98.95

through the WHU-CD dataset, and use F1, IoU, Recall, OA, and Precision to compare the effects. The specific experimental results are shown in Table VIII. By observing Table VIII, we can find that when the value of α is 0.5, that is, the high-frequency and low-frequency branches in HiLo adopt the same number of attention heads, the effect modeling ability of the model is the strongest, and the relationship between the simulated features can be more realistic. We believe that the information carried by

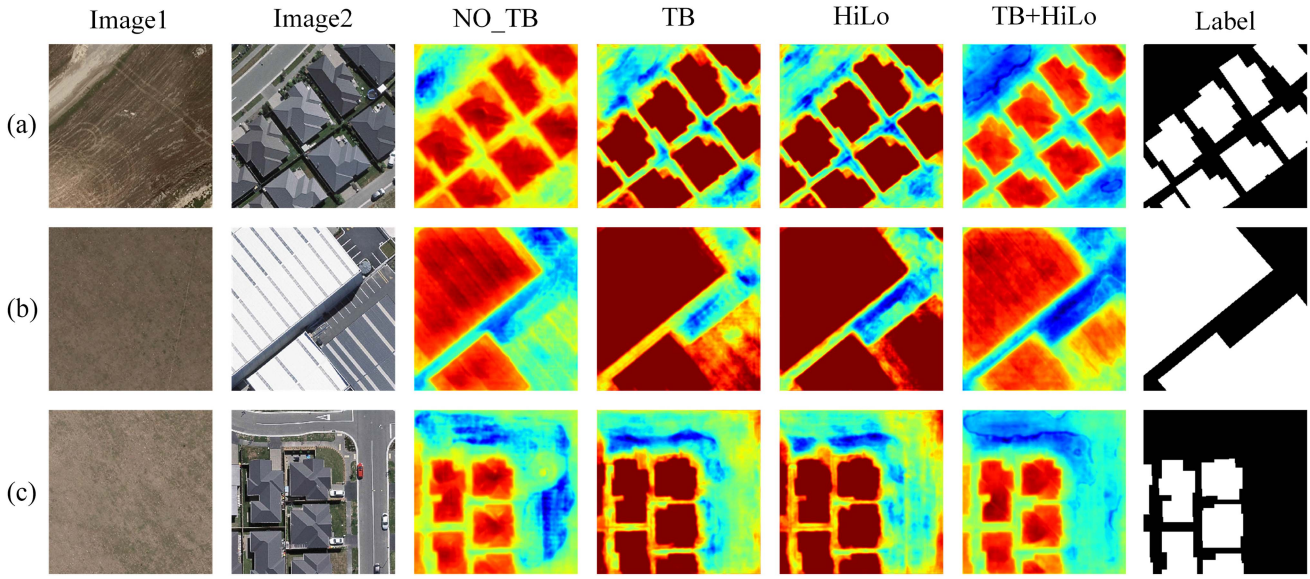


Fig. 10. Visualization of attention maps during testing. (a), (b), and (c) respectively represent three different sets of bi-temporal images selected from the WHU-CD test set for visualizing the effect of the ablation module.

TABLE IX
EXPERIMENTAL ANALYSIS OF THE INFLUENCE OF TOKENIZER ENCODER AND TOKENIZER DECODER LAYERS ON THE ESACD MODEL ON THE WHU-CD DATASET

N	M	Precision	Recall	F1	IoU	OA
1	8	90.72	91.50	91.11	83.67	99.14
2	6	92.76	88.17	90.41	82.49	99.10
4	4	89.34	90.90	90.11	82.00	99.04
6	2	88.76	88.81	88.78	79.83	98.92
8	1	90.79	90.47	90.63	82.86	82.86

low-frequency and high-frequency in feature maps is mined and expressed, which strengthens the model's extraction of global structural information and local fine features, and strengthens the model's expressive ability. At the same time, we choose $\alpha = 0.5$ as the optimal value, and we can infer that the low-frequency information in the image is as important as the feature content contained in the high-frequency information.

Next, we analyze the parameters of the layers N and M of the self-attention mechanism in the Tokenizer encoder and the Tokenizer decoder. We choose (1, 8), (2, 6), (4, 4), (6, 2), and (8, 1) for N and M, respectively, to analyze the model effect, and the experimental results are shown in Table IX.

By observing Table IX, we can find that when the values of N and M are 1 and 8, our model can achieve the best performance. We think that our model needs more Tokenizer decoders to project tokens into pixel-level features. Compared with the Tokenizer encoder for context modeling, this requires more powerful feature representation capabilities to obtain more detailed semantic features.

F. Visualization

In order to understand the role of Tokenizer attention and HiLo attention more intuitively, we visualize the effect diagrams of ESACD (TB), ESACD (NO_{TA}), ESACD (NO_{HiLo}), and ESACD (TA+HiLo), as shown in Fig. 10. Higher values of attention are shown in red and lower values in blue. In the ESACD (HiLo) column and ESACD (TA) column, we can see that the existence of HiLo attention and Tokenizer attention makes the model pay more attention to the main body of the building, and both of them can extract richer semantic information for large buildings and small buildings for efficient modeling. In the TA+HiLo column, we can observe that the combination of HiLo attention and Tokenizer attention makes the model's segmentation of the boundary of the changing building body more refined and has a clear boundary. I think this is a proof of the effectiveness of our dual attention. From left to right, the model's concerns and nonconcerns are more clearly defined. Because the model pays less attention to nonbuilding areas, there are more areas of blue pixels. At the same time, the change map predicted by the ESACD (TA+HiLo) column is closer to label.

To further improve the persuasion of the high- and low-frequency attention mechanism, we visualize the frequency information of feature maps in the high- and low-frequency branches. As shown in Fig. 11, the Hi-Fi group and the Lo-Fi group clearly show the frequency information of the feature maps obtained by the high- and low-frequency branches.

G. Discussion

Our model outperforms other state-of-the-art methods on WHU-CD dataset and Google-CD dataset. However, in the face of remote sensing image datasets with high aerial photography heights, such as Onera satellite change detection, our ESACD does not perform well. We believe that ESACD lacks the ability to recognize different building types in the same images. In the face of architectural diversity and relatively low image quality, we believe that it is necessary to strengthen the ability

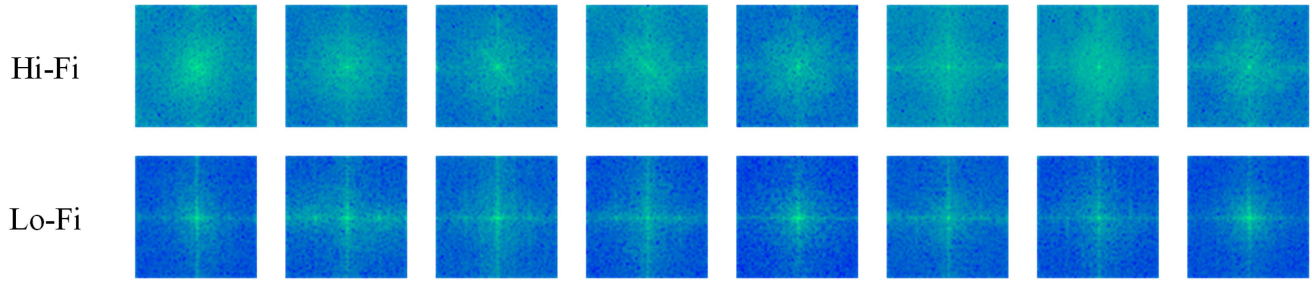


Fig. 11. Visualization of high- and low-frequency attention.

to capture high-frequency feature information in HiLo attention and strengthen the learning of detailed features. And continue to improve the model from the direction of the model architecture. Due to the relatively lightweight characteristics of our model, we may choose to deploy and apply the model to the mobile platform to optimize the running speed and achieve fast and efficient remote sensing image change detection as much as possible. At the same time, the model needs to enhance the feature expression ability to deal with the situation of architectural diversity. In addition, in this work, we process the features extracted by dual attention in a pixel-by-pixel way, which may not be the most effective fusion method. In future work, we will focus on the research and analysis of the above two issues to improve the effect of our model.

V. Conclusion

In this work, we propose a neural network ESACD based on the combination of CNN and transformer and utilizing the In this work, we propose a lightweight change detection model for remote sensing building images. We use dynamic context and static context to replace the ordinary convolutional layer to model the context relationship, and the dynamic context performs self-attention operation under the guidance of the static context. Subsequently, in order to enhance the learning ability of the model, we applied dual attention to simulate the feature relationship for the low-frequency information and high-frequency information in the images and the features of interest to strengthen the effect of the model. In HiLo attention, the low-frequency and high-frequency semantic features in the images are mined while maintaining low complexity, and the two are fused. Tokenizer attention extracts the features of interest to the model and encodes them into pixel-level features. Dual attention greatly enhances the modeling ability of the model, showing excellent performance. Finally, we fuse the semantic features of different resolutions, on the basis of enhancing the fluidity of information, so that the detailed features can be fully integrated. After a large number of experiments, it has been proved that our model has a surprising effect compared with other excellent comparison methods in the case of a small difference in the number of parameters.

References

[1] K. L. de Jong and A. S. Bosman, "Unsupervised change detection in satellite images using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[2] D. Zhang and G. Zhou, "Estimation of soil moisture from optical and thermal remote sensing: A review," *Sensors*, vol. 16, no. 8, 2016, Art. no. 1308.

[3] F. Wang and Y. J. Xu, "Comparison of remote sensing change detection techniques for assessing hurricane damage to forests," *Environ. Monit. Assessment*, vol. 162, no. 1, pp. 311–326, 2010.

[4] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

[5] C. Marin, F. Bovolo, and L. Bruzzone, "Building change detection in multitemporal very high resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2664–2682, May 2015.

[6] Y. Gao, F. Gao, J. Dong, and S. Wang, "Change detection from synthetic aperture radar images based on channel weighting-based deep cascade network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 11, pp. 4517–4529, Nov. 2019.

[7] K. Rokni, A. Ahmad, A. Selamat, and S. Hazini, "Water feature extraction and change detection using multitemporal Landsat imagery," *Remote Sens.*, vol. 6, no. 5, pp. 4173–4189, 2014.

[8] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4227–4240, Nov. 2021.

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

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[11] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[12] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[13] C. Wu, F. Wu, T. Qi, Y. Huang, and X. Xie, "Fastformer: Additive attention can be all you need," 2021, *arXiv:2108.09084*.

[14] J. W. Cooley, P. A. Lewis, and P. D. Welch, "The fast Fourier transform and its applications," *IEEE Trans. Educ.*, vol. TE-12, no. 1, pp. 27–34, Mar. 1969.

[15] G. Deng and L. Cahill, "An adaptive Gaussian filter for noise reduction and edge detection," in *Proc. IEEE Conf. Rec. Nucl. Sci. Symp. Med. Imag. Conf.*, 1993, pp. 1615–1619.

[16] M. Xiao et al., "Invertible image rescaling," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 126–144.

[17] J. Huang, D. Guan, A. Xiao, and S. Lu, "RDA: Robust domain adaptation via Fourier adversarial attacking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8988–8999.

[18] Y. Zhou, W. Deng, T. Tong, and Q. Gao, "Guided frequency separation network for real-world super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 428–429.

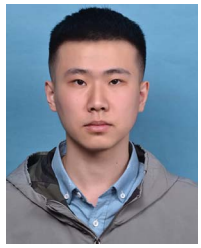
[19] M. Fritsche, S. Gu, and R. Timofte, "Frequency separation for real-world super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 3599–3608.

[20] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1740–1749.

[21] W. Chen, J. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing convolutional neural networks in the frequency domain," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1475–1484.

[22] Z. Pan, J. Cai, and B. Zhuang, "Fast vision transformers with HiLo attention," 2022, *arXiv:2205.13213*.

[23] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023.
[24] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2021, Art. no. 5607514.
[25] J. Zhang and Y. Zhang, "Remote sensing research issues of the national land use change program of China," *ISPRS J. Photogrammetry Remote Sens.*, vol. 62, no. 6, pp. 461–472, 2007.
[26] 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.
[27] B. Hou, Y. Wang, and Q. Liu, "Change detection based on deep features and low rank," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2418–2422, Dec. 2017.
[28] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1382.
[29] Y. Lei, X. Liu, J. Shi, C. Lei, and J. Wang, "Multiscale superpixel segmentation with deep features for change detection," *IEEE Access*, vol. 7, pp. 36600–36616, 2019.
[30] N. Lv, C. Chen, T. Qiu, and A. K. Sangaiah, "Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images," *IEEE Trans. Ind. Inform.*, vol. 14, no. 12, pp. 5530–5538, Dec. 2018.
[31] A. M. El Amin, Q. Liu, and Y. Wang, "Zoom out CNNs features for optical remote sensing change detection," in *Proc. 2nd Int. Conf. Image, Vision, Comput.*, 2017, pp. 812–817.
[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[33] X. Wang et al., "A high-resolution feature difference attention network for the application of building change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102950.
[34] 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.
[35] H. Zhang, M. Gong, P. Zhang, L. Su, and J. Shi, "Feature-level change detection using deep representation and feature change analysis for multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1666–1670, Nov. 2016.
[36] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised multiple-change detection in VHR multisensor images via deep-learning based adaptation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5033–5036.
[37] H. Zhang, M. Lin, G. Yang, and L. Zhang, "ESCNet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 28–42, Jan. 2023.
[38] 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.
[39] 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.
[40] X. Su, J. Li, and Z. Hua, "Transformer-based regression network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–23, Feb. 2022.
[41] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
[42] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," 2020, *arXiv:2004.11886*.
[43] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "DeLighT: Very deep and light-weight transformer," 2020, *arXiv:2008.00623*.
[44] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," *IEEE Int. Geosci. Remote Sens. Symp.*, pp. 207–210, 2022.
[45] M. Liu, Q. Shi, Z. Chai, and J. Li, "PA-Former: Learning prior-aware transformer for remote sensing building change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Aug. 2022, Art. no. 6515305.
[46] Z. Wang, Y. Zhang, L. Luo, and N. Wang, "TransCD: Scene change detection via transformer-based architecture," *Opt. Exp.*, vol. 29, no. 25, pp. 41409–41427, 2021.
[47] Z. Li, J. Li, F. Zhang, and L. Fan, "CADUI: Cross attention-based depth unfolding iteration network for pan-sharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402420.
[48] T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, and A. Plaza, "A Siamese network based U-Net for change detection in high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2357–2369, Mar. 2022.
[49] H. Zheng et al., "HFA-Net: High frequency attention Siamese network for building change detection in VHR remote sensing images," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108717.
[50] 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, 2021.
[51] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.
[52] 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.
[53] 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.



Shike Liang received the B.S. degree in computer science and technology from the School of Information Engineering, Shandong Youth University of Political Science, Jinan, China, in 2021. He is currently working toward the master's degree in electronic information with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai, China.

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

Zhen Hua received the B.S. and M.S. degrees in electrical automation from Taiyuan University of Technology, Taiyuan, China, in 1989 and 1992, respectively, and the Ph.D. degree in electronic information engineering from China University of Mining and Technology, Beijing, China, in 2008.

She is currently a Professor with Shandong Technology and Business University, Yantai, China. Her research interests include computer-aided geometric design, information visualization, virtual reality, and image processing.

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

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.