GateFormer: Gate Attention UNet With Transformer for Change Detection of Remote Sensing Images

Li-Li Li, Zhi-Hui You, Si-Bao Chen, *Member, IEEE*, Li-Li Huang, Jin Tang, and Bin Luo, *Senior Member, IEEE*

***Abstract*—Extraction of global context information plays a major role in change detection (CD) of remote sensing (RS) images. However, the majority of methods now depend on convolutional neural networks, which are difficult to obtain complete context information due to the limitation of local convolution operation. This study proposes a novel gate attention U-shaped network with transformer for CD of RS images. GateFormer consists of an encoder with transformer-based Siamese network. First, we propose a gate attention mechanism, which filters the low-level information by guiding high-level features and focuses on activation of relevant knowledge instead of allowing all to pass. In addition, space pooling module in generator extracts more spatial features from pixel level to suppress the generation of noises. Finally, in order to increase the CD accuracy of small-scale ground objects, we design a feature downsampling module to minimize the loss of detailed information and compress more small-scale features in feature downsampling of transformer. The efficiency of our suggested approach has been verified by experiments on three RS CD datasets.**

***Index Terms*—Change detection (CD), gate attention, remote-sensing (RS) image, transformer, U-shaped network (UNet).**

I. Introduction

IN ORDER to identify changes in the area of interest, change detection (CD) in remote sensing (RS) analyzes two (or more) images from the same region at different periods [1]. RS image CD has gradually developed as one of the key research directions in the field of RS due to the rapid development of RS technology in recent years. At present, RS image CD has been applied in many real-world scenario applications, such as land resource management [2], disaster assessment [3], urban expansion [4], [5], and other fields.

Traditional CD methods can be roughly divided into two categories: pixel-based CD methods and object-based CD methods. Pixel-based CD methods usually generate a difference map to compare the spectral information of the bi-temporal image, and finally obtain a threshold or clustering method for division. It is worth noting that pixel-based CD methods, include image differencing [6], image ratio [7], regression analysis [8], change vector analysis [9], and principal component analysis [10]. But pixel-based methods ignore the connectivity information of the class context. In addition, their performance largely depends on the decision function and threshold setting. Later methods are usually based on and the object-based method is to introduce the idea of class segmentation to extract information from the segmented images, so as to identify the changes between the bitemporal images. Zhang et al. [11] used the incremental segmentation method to segment the image to construct the object-based feature space of bitemporal, and then calculated the change index of each object through the cosine law to identify the changed object. However, the segmentation performance of object-based CD methods also has some shortcomings, and there may be undersegmentation and oversegmentation errors. High-resolution images present complex textures and details, which bring new challenges to the CD task.

Recently, CD using deep learning techniques has shown results worthy of celebration due to their ability to extract intricate features from images. Convolutional neural networks (CNNs) are being successfully applied in various kinds of RS image fields, such as object detection [12], [13], [14], image classification [15], [16], [17], semantic segmentation [18], [19], and so on. The encoder–decoder structure in this method naturally completes the end-to-end training, which not only shows excellent segmentation performance in semantic segmentation and also displays a significant effect in CD field. For example, the U-shaped network (UNet) [20] encoder captures the feature information of different levels, and realizes the fusion of information carried in the encoder and decoder by skipping connections to continuously restore the image resolution. The integration of spatial features at different levels greatly enhances network performance.

At the same time, many excellent CD algorithms have been proposed. An example is depth change vector analysis [21], which leveraged multiple layers of CNNs to model the spatial relationships between neighboring pixels and complex objects. Zhan et al. [22] introduced the first CD work of Siamese convolutional networks, which could realize parallel processing of bitemporal images. From then on, there were many methods in CD based on Siamese structures rather than simply fusing images as input. Three full CNNs were developed by

Manuscript received 18 August 2023; revised 10 October 2023 and 6 November 2023; accepted 14 November 2023. Date of publication 21 November 2023; date of current version 6 December 2023. This work was supported in part by the NSFC Key Project of International (Regional) Cooperation and Exchanges under Grant 61860206004, in part by the NSFC Key Project of Joint Fund for Enterprise Innovation and Development under Grant U20B2068, and in part by the National Natural Science Foundation of China under Grant 61976004. *(Corresponding authors: Si-Bao Chen; Li-Li Huang.)*

The authors are with the MOE Key Lab of ICSP, IMIS Lab of Anhui Province, Anhui Provincial Key Lab of Multimodal Cognitive Computation, Zenmorn-AHU AI Joint Lab, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: 627154942@qq.com; 1574583514@qq.com; sbchen@ahu.edu.cn; hill_ahu@ahu.edu.cn; tj@ahu.edu.cn; luobin@ahu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3335281

Daudt et al. [23] to address the issue of CD. FC-EF connected bi-temporal images before entering the network, while FC-Siam-conc and FC-Siam-diff leveraged a Siamese structure, which could directly share network weights. However, the fusion of bitemporal features in the encoder fails to fully explore the correspondence of bitemporal, resulting in the lack of multiscale semantic information. Liu et al. [24] designed a bitemporal feature fusion module to enrich the multiscale and improve the correspondence by filtering the background noise. Song et al. [25] introduced spatial and channel attention to better distinguish between foreground and background in order to enhance the capacity to recognize changes. The application of attention mechanism had also brought remarkable results for RS image processing [26], [27], [28], [29].

Although some of the methods mentioned above have been successful, most rely on convolution operations and the RS image data of CD has more abundant spatial and shape features. For example, due to light angle, shadow, seasonal change, reflection of ground objects, and other reasons, the same objects may present different characteristics. In addition, branch occlusion can also lead to false changes. The model based on CNNs [30], [31], [32] only carries out convolution on the regular rectangular region, which means that the network using CNN can hardly collect the complete shape details and contextual inference of the object. However, more global features and detailed spatial features are needed to support the identification of changing regions.

Transformer [33] was originally proposed for natural language processing. Subsequently, it has been attracted a lot of attention in the field of computer vision, and has opened up new research ideas for researchers. Vision transformer (ViT) [34] was the first application of transformer methods to image classification, it introduced self-attention module for modeling long-range information by pairwise interaction between each patch block. We could see that transformer used its powerful global processing capabilities to perform as well as CNNs on many large datasets, or even better. Currently, the transformer-based models have greatly improved at image classification [35], object detection [36], semantic segmentation [37], [38], road extraction [39], and video captioning [40], [41]. But its CD potential on RS images remains to be developed.

Inspired by the achievement of transformer, we introduce the GateFormer, a novel network structure for CD, which integrates transformer into the UNet [20] framework and builds a pyramid of multilevel, multiscale feature objects. Transformer is used as an encoder in our method to extract features, establish distant information, and create more useful global features. GateFormer can thus capture as much global information in space–time as possible without being limited by convolution operation. At the same time, the designed Siamese structure is also more suitable for CD bi-temporal images extraction. Each unit block in the encoder primarily uses transformer to extract features. In addition, we take the Unet architecture as the backbone and use gate attention mechanism (GAM) before skipping connections, focusing on activating relevant regions. Next, inspired by the strip pooling module [42], it deploys a long strip-shaped pooling kernel along different spatial dimensions, enabling the network to effectively simulate long-distance dependencies. Space pooling module (SPM) explores the spatial correlation of global features from two different directional dimensions and reduces the generation of noisy information. Feature downsampling module (FDM) is used to increase the accuracy of small-scale detection. and avoid missing feature information in downsampling. The following are the primary contributions of the proposed work.

1) With the UNet structure as the backbone, the skip connection of the GAM is constructed, which focuses on the activation of relevant region.
2) The SPM is proposed, which dedicates to extracting pixel-level features in the spatial dimension, thus compensating for the transformer's regret in global feature extraction.
3) In the transformer encoder, the FDM is designed to alleviate the neglect of small objects in the downsampling process and take care of the extraction of global information.

II. Related Work

A. CNN-Based CD Methods

Due to the powerful discrimination ability of CNN to effectively distinguish real changes from complex irrelevant changes, various CNN-based CD methods have been proposed and achieved good achievements. Current CD methods almost all of them are based on convolution operation. Zhang et al. [43] brought in deep features using a deep belief network, then represented change areas by the difference of blocks. Lei et al. [44] presented a UNet-structured network that used a pyramid pooling module to obtain multiscale change information. Peng et al. [45] proposed UNet++ network with multiside output fusion strategy. In order to fully utilize the available spatial information, Daudt et al. [23] proposed three fully convolutional network architectures. They are EF, FC-Sim, and FC-Siam-diff, including one cascading operation and two Siamese neural networks. However, convolution kernel only deals with local domain in time and space, and has defects in capturing long-distance global interaction, this makes it challenging to adjust RS images for size, shape, and position changes. Therefore, attention mechanisms (channel attention, spatial attention, and self-attention) are usually used to solve this problem and further increase the intrinsic semantic relationship between image pairs. For example, DASNet [46] introduced a dual attention module to learn channel and spatial information features in the CD task. Fang et al. [47] proposed an integrated channel attention module to fuse feature information of different levels, and there are a self-attention mechanism based on STANet [48] and a deeply supervised attention metric-based network [49]. Due to the lack of supervision in the middle layer (or lower layer) during the training of deep convolutional networks, the modeling of semantic relationships is weakened. Therefore, Zhang et al. [27] proposed a deep supervised image fusion network (IFN) to input the extracted deep features into a deep supervised differential recognition network to enhance the effectiveness of CD.

Although the methods using CNN have contributed to some progress in the direction of CD, the convolution operation still unable to effectively extract long-distance global features

and cannot adapt to scenarios that are complex, susceptible to noise, or have very different training sets, which limits the improvement of CD accuracy to a certain extent. This article is different from these previous methods, we try to use transformer to establish long-distance global interaction.

B. Transformer-Based CD Methods

The ViT has recently attracted the attention of many researchers. Vaswani et al. [33] applied pure transformer to image classification through continuous adjustment and processing, and finally achieved good results. Subsequently, the hierarchical Swin transformer that Liu et al. [50] proposed shifting windows to restrict self-attention computation to nonoverlapping local windows and allow connections across windows. Due to transformer-based methods exhibit comparable or even better performance than convolutions, it corresponds to various image processing task, including image classification, segmentation, image generation, and object detection.

When it comes to CD, transformer-based methods are starting to become more. Yan et al. [51] proposed a framework called the full transformer network (FTN) that included multilevel features in a pyramid structure, which is beneficial to extract features from a global view. To solve the CD challenge. Zhang et al. [52] used a pure Swin transformer network with a Siamese U-shaped structure, both encoder and decoder was based on hierarchical Swin transformer. Zheng et al. [53] designed a deep multitask METD architecture for semantic CD and Wang et al. [54] incorporated the twin vision transformer (SVIT) into the feature difference framework for CD. In order to take advantage of transformer and CNN, Chen et al. [55] developed the BIT-CD model by fusing the CNN and transformer architectures, which models the context in the spatial-temporal domain. Bandara and Patel [56] used a multilayer perception machine (MLP) as a decoder and a hierarchical structure of transformer encoders to produce multiscale remote features. Feng et al. [57] proposed an intrascale cross interaction and scale feature fusion network (ICIF-Net), which used CNN and transformer to extract local and global features, respectively, and invoked mask aggregation and spatial alignment (SA) for feature fusion at different scales.

However, these methods ignore the importance of high-resolution latent information in the feature extraction process, resulting in a lot of noise background, which is not conducive to fully paying attention to small changes in complex scenes. Our proposed GateFormer architecture makes full use of the multiscale information and the internal relationship of the bi-temporal image through hierarchical transformer and fusion module. At the same time, we introduces gate attention to guide high-level information to low-level information for better depiction of details and boundary region features, as well as precise localization of small-scale objects.

III. Methodology

In this section, we first introduce the general structure of GateFormer and describe the structural principle of transformer feature extraction. Then three crucial modules in GateFormer are introduced, they are GAM, SPM, and FDM.

Algorithm 1: Implementation Steps of the GateFormer Model.

Input: Bitemporal remote sensing images $\mathrm{I} = \{(I_1, I_2)\}$
Output: A prediction change mask Z_1

// Get different scale information by SPM and FDM
for $i\ in\{1,2,3\}$ **do**
  for $j\ in\{1,2\}$ **do**
    if $i = 1$ **then**
      $X_{i,j} = SPM(FDM(I_j))$
    else [i=2 or 3]
      $X_{i,j} = SPM(FDM(X_{i-1,j}))$
    end if
  end for
  $F_i = Fusion\ Module(X_{i,j})$
end for
// Concatenate the output of the first three stages and input the results again into the final FDM and SPM
$F = Concat(F_1, F_2, F_3)$
$F_4 = SPM(FDM(F))$
// Upsampling in the decoder
for $i\ in\{4,3,2\}$ **do**
  if $i = 4$ **then**
    $Z_i = GAM(F_{i-1}, F_i)$
  else [i=3 or 2]
    $Z_i = GAM(F_i, Z_{i+1})$
  end if
end for
// Get the map through through the prediction head
$Z_1 = Prediction\ Head(Z_2)$

A. Network Structure

Fig. 1 depicts the overall design of our GateFormer. As an integration between the transformer and UNet. The encoder and decoder are connected by skipping connection layers in our GateFormer, which also uses a feature fusion module to fuse the bi-temporal hierarchical data. The decoder receives the output from each layer of the feature fusion module directly. The output size of each layer transformer block decreases as the number of channels increase, as seen in Fig. 1. In addition, to further enhance the performance of CD, we also design GAM, SPM, and FDM.

The GateFormer's in detail operation is illustrated in Algorithm 1.

For the initial bitemporal RS images $I \in \mathbb{R}^{H\times W\times 3}$. To symbolize the "tokens" of sequence data, ViT separates the image data into flat, uniform, and nonoverlapping patches. After flattening and projecting these "tokens" to dimension C_1, the linear embedding layer puts in the encoder stacked by the transformer blocks. There are four feature extraction steps in the encoder. We define the output of each layer of prechange or post change images as F_n, where $n = 1, 2, 3$, and 4. In particular, our SPM module can easily establish pixel level information exchange, thus, making up for the limitation of pure transformer based on limited windows, and can extract more detailed features. In addition, in order to pay more attention to the region of interest, we use GAM to achieve feature transfer from the encoder to the decoder. In this

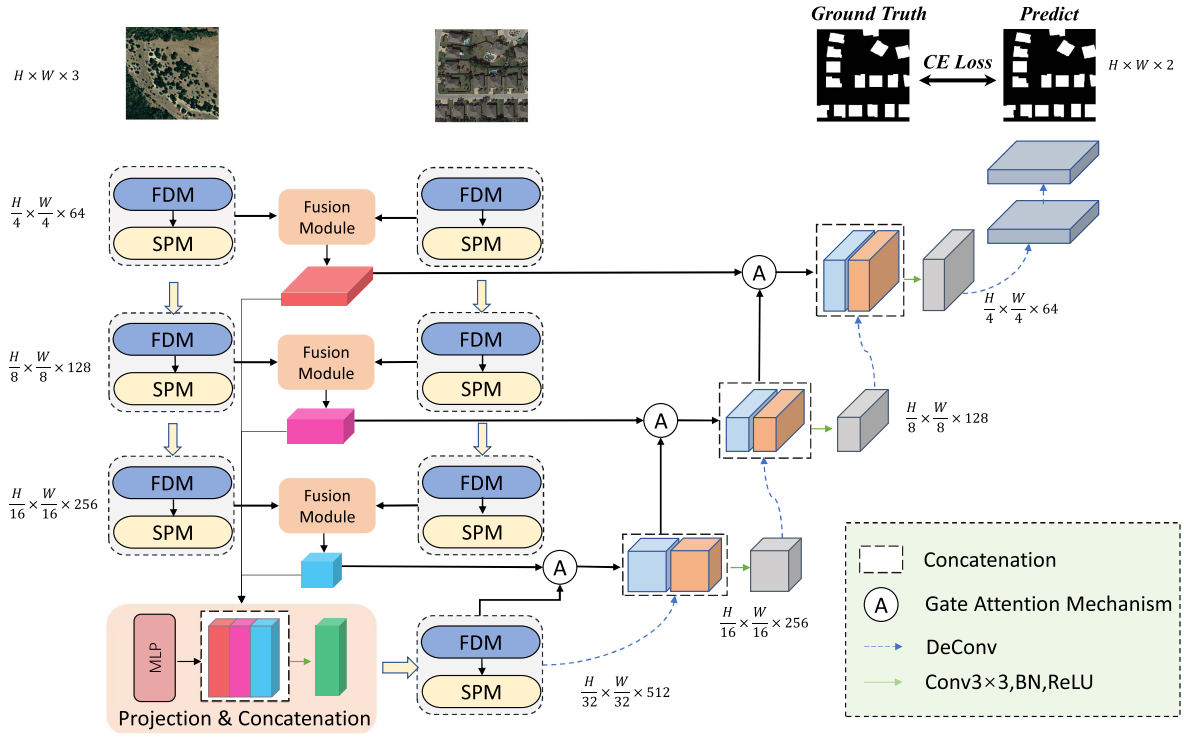


Fig. 1. Framework overview of GateFormer. A linear layer is used to project the bi-temporal images into embedding space after dividing them into a number of non-overlapping patches. A position embedding is included to the sequence before it is fed into the encoder. GateFormer contains three important modules: GAM, SPM, and FDM.

process, our FDM can reduce the ignorance of small-scale object details. The dimensions of stage n are $2n-1C_1$, and the output resolution is $(H/(2n+1)) \times (W/(2n+1))$. In the decoder of each layer, the n block's output feature map may be expressed as $F_n \in \mathbb{R}^{(H/(2^{n+1}))\times(W/(2^{n+1}))\times 2^{n-1}C_2}$. Here, $C_2 = 64$. Then, in order to obtain F_{fus}, the output of the transformer at the right stage will then enter the feature fusion module. It is worth noting that the output of the first three layers of the fusion module is projected and concatenated into $\mathbb{R}^{(H/16)\times(W/16)\times 256}$, and fed into the final separate transformer decoding block for encoding. At this time, the encoding content contains the global context of all high-level semantic concepts and features of different levels, with the aim to better finish the recovery of details.

After the mentioned four encoding phases, we obtain the $\mathbb{R}^{(H/32)\times(W/32)\times 512}$ feature size, and put it in a 2×2 deconvolution layer to increase the resolution. Here, GateFormer uses skip connections to connect the upsampling results in the encoder and decoder for each pass gate attention. At the same time, a 3×3 convolution layer is used to reduce the number of channels. The process mentioned above is run three times, progressively expanding feature Z to $\mathbb{R}^{(H/4)\times(W/4)\times 64}$. We employ two 3×3 convolutional layers and linear interpolation upsampling to feature Z to produce the final projected binary classification results.

B. Transformer Block

Compared with the CNN, ViT [33] relies on excellent context capabilities to achieve very excellent performance on multiple benchmarks. Transformer block consists of n-layer multihead self-attention (MSA) and multilayer perceptron (MLP) blocks. Each MSA, MLP is connected by a layer normalization (LN) module in front, and a residual connection behind. Formally

$$z'_i = \text{MSA}(\text{LN}(z^{i-1})) + z^{i-1}, i = 1 \ldots L \tag{1}$$

$$z_i = \text{MLP}(\text{LN}(z'_i)) + z'_i, i = 1 \ldots L \tag{2}$$

where i is the identifier for the intermediate block, L is the number of transformer layers, $\text{LN}(\cdot)$ stands for layer normalization, and MLP consists of two linear layers with GELU activation functions.

Among them, multihead attention mechanism is the transformer's primary component. Transformer can acquire information from different head parts and combine it using MSA, which also gives it the ability to understand context. In the initial work, the self-attention is described as

$$\text{SA}(z) = \sigma\left(\frac{(QK^{\top})}{\sqrt{d}}\right)V \tag{3}$$

where Q, K, and V denote query, key, and value, respectively, and d is the channel dimension of the triple, the Softmax function applied to the channel dimension is shown by the symbol σ.

MSA is composed of several parallel independent attention heads, the advantage of MSA is that it can simultaneously handle data from various representation subspaces located in many locations. Formally

$$\text{MSA}(z) = [\text{SA}_1(z); \text{SA}_2(z); \ldots; \text{SA}_h(z)]w_{\text{msa}} \tag{4}$$

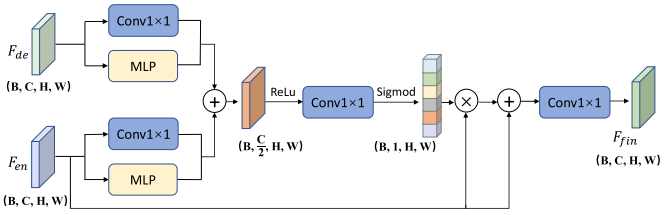


Fig. 2. Structure of GAM. F_{en} and F_{de} are from the corresponding encoder and decoder, respectively.

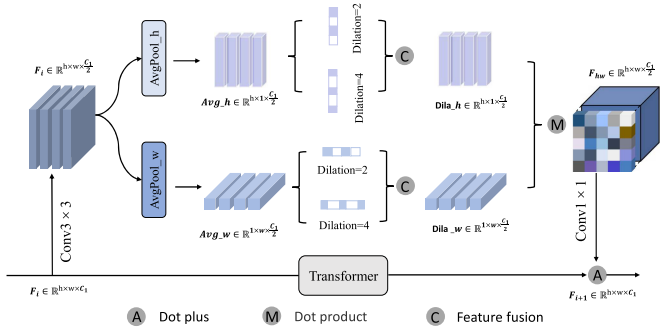


Fig. 3. Structure of SPM aggregates information in vertical and horizontal spatial directions, respectively.

where h is the number of attention heads and $w_{\text{msa}} \in R^{hd\times C}$ is the linear projection matrices.

Before supplying the acquired patches to the transformer layers, we add a 1-D learnable positional embedding to maintain their spatial information. Such position information may better play to the characteristics of parallel input. In other words, remember the relative position between patches.

C. Gate Attention Mechanism

Accurate change region detection needs deeper networks that learn more complex features from the data. However, more spatial information will be lost as the network gets deeper, decreasing the precision of CD. In order to transfer contextual and spatial data between encoder and decoder, UNet applies skip connections, and in this way alleviates the loss of information during downsampling to some extent. GateFormer uses structure of UNet to transfer information between the encoder and decoder, the gate attention module is used for the activation of relevant features in the decoder. This mechanism can suppress the feature activation of irrelevant regions in the images and reduce the detection error. The gate attention connects the fusion module in the encoder and its corresponding decoder, as shown in Fig. 1.

GAM is conveyed by two branches: F_{en} from the fusion module of the corresponding encoder, which contains all the contextual and spatial information of the bitemporal images in the corresponding layer, and F_{de} from the decoder layer below. Here, the size and number of channels of the input of these two branches are consistent. Note that the decoder receives additional access to the gate's output for cascading. Fig. 2 displays the representation of the gate attention applied in the model we propose. The main purpose of 1×1 convolution is to reduce the number of channels to simplify computation, but it will lose some important feature details while reducing the number of channels. Adding parallel MLP can better capture the characteristics of the input data and carry more useful information to complement the surrounding information, so as to reduce the dimension information and maximize the feature to avoid the loss of feature information, which is more suitable for future CD tasks. Then, the two processed input features are combined, and the sigmod function is used to obtain the probability weight coefficient, denoted by P_g. The above operation can be expressed as follows:

$$P_g = \delta(f([\text{MLP}(g); f(g)] + [\text{MLP}(x); f(x)])) \tag{5}$$

where δ stands for the sigmod function and a convolution operation with a filter size of 1×1 is represented by $f(\cdot)$. Finally, We multiply the channel correlation P_g by the feature F_{en} from the encoder to obtain a more refined feature F_{fin} to achieve the role of using high-level information to guide the analysis of low-level data. As shown in the following equation:

$$F_{\text{fin}} = f(F_{\text{en}} \odot P_g + F_{\text{en}}) \tag{6}$$

where $\odot$ represents element-level multiplication.

D. Space Pooling Module

Since transformer performs self-attention computation between token vectors mapped by input partitioned patches. It effectively reduces the memory overhead compared with the interaction between each pixel. However, transformer's global modeling capabilities are somewhat diminished by this strategy. Moreover, in CD datasets, there may be identical intervals, trees, or light occlusions that lead to wrong judgments, which requires some spatial information to change this phenomenon. Therefore, in addition to the main branch of transformer, another branch is introduced: spatial pooling module (SPM), which brings attention to two spatial dimensions: horizontal features and vertical features. Among them, the extraction of horizontal information can better capture the local and details of the data, so as to better express the characteristics of the input data. The purpose of vertical feature extraction is to better combine and classify the data that has or has not changed, with the aim to more effectively capture the overall data and abstract concept. Thus, horizontal and vertical are introduced together to consider the relationship between pixels, taking care of not only local and detailed features but also global concepts. In addition, it compensates for the fact that the transformer only focuses on each patch token.

As shown in Fig. 3, for stage n, we first reshape the input feature $F \in \mathbb{R}^{(hw)\times c_1}$ to $F \in \mathbb{R}^{h\times w\times c_1}$. Here, $c_1 = 2^{n-1}C_1$, $h = (H/(2n+1))$, and $w = (W/(2n+1))$. Before the vertical and horizontal operations, the feature F is inputted into a 3×3 convolutional layer, which the operation reduces the number of channels to $c_1/2$ to reduce the computational cost. In the horizontal direction, the $1 \times h$ pooling operation is first used to turn the feature map into a feature representation with shape

$h \times 1$, and on this basis, a set of parallel dilated convolutions are used to extract features from the feature map. We set the dilation rate to $[2, 4]$. This operation can further increase the horizontal receptive field without additional parameters and perceive objects in the horizontal direction from multiple scales at the same time. Then, the feature maps generated by these parallel branches are fused by concatenate feature fusion method. The total tensor in the horizontal direction is denoted by Avg_$h = \mathbb{R}^{h\times 1\times c_1/2}$. The operation in the vertical direction is similar to that in the horizontal direction, and the tensor in the vertical direction is denoted by Avg_$w = \mathbb{R}^{1\times w\times c_1/2}$. In the vertical direction, the $1 \times w$ pooling operation is used to turn the feature map into a $w \times 1$ feature representation, and the subsequent operations are similar to the horizontal direction. The following formula is mentioned for calculating the elements in each direction:

$$\text{Avg\_}h = \frac{1}{w}\sum_{j=0}^{w-1}\hat{z}^k(i,k) \tag{7}$$

$$\text{Avg\_}w = \frac{1}{h}\sum_{i=0}^{h-1}\hat{z}^k(i,k) \tag{8}$$

where the indexes for the channel, the horizontal direction, and the vertical direction are i, j, and k, respectively. In this case, $0 \le i < h$, $0 \le j < w$, and $0 \le k < c_1/2$. $f_d(\cdot)$ is the dilated convolution layer with batch normalization and the GELU activation function. Avg_h and Avg_w extract the weights of pixel information in different dimensional spaces. The feature $\hat{z}^k = f_d(z)$.

After obtaining the horizontal and vertical features, they are multiplied for feature combination, and then the combined features are fed into a 1×1 convolution for feature adjustment. Finally, this output is superimposed with the output of the transformer feature processing to obtain our result. Feature $F \in \mathbb{R}^{h\times w\times c_1}$ can be expressed as follows:

$$F_{hw} = f_\xi(\text{Avg\_}h(F_i)) \times f_\xi(\text{Avg\_}w(F_i)) \tag{9}$$

where f_ξ represents the convolution parallel operation with expansion rates of 2 and 4, respectively, and the operation results F_{hw} is the feature information extracted by SPM.

E. Feature Downsampling Module

Downsampling can not only reduce the amount of calculation to prevent overfitting, but also increase the received field. More global information can be learned by downsampling the convolution kernel. Most transformer downsampling methods are to merge adjacent blocks or map directly, but such methods are too direct, which will lead to the loss of many small objects in the CD dataset. Therefore, we designed FDM in the hierarchical transformer downsampling to compensate for the negligence of small objects.

Specifically, Fig. 4 depicts the two branches of the FDM. One is the block with expansion convolution, which by expanding the receptive field of convolution captures the properties and structural details of small-scale objects. The first 1×1 convolutional layer in the bottleneck block increases dimension. Detailed structural information is obtained using the middle

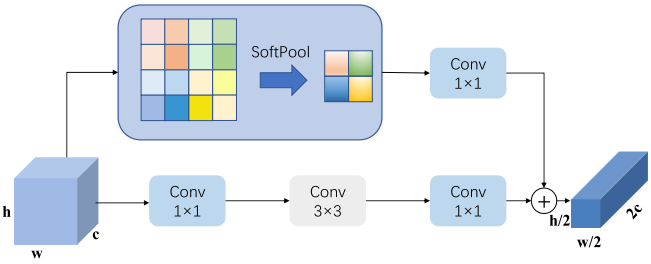


Fig. 4. Depiction of FDM. FDM reduces resolution through two different branches, soft-pool operation and convolutional layer to reduce the resolution and obtain finer features.

3×3 dilated convolutional layer, and feature scale is reduced using the final 1×1 convolutional layer. The output of this branch $F_1 \in \mathbb{R}^{h/2\times w/2\times 2c_1}$. Another is to reduce the feature scale through soft-pool [58] operation to obtain more detailed features. For the purpose of preserving the fundamental properties of the input, soft-pool accumulates activation in an exponentially weighted manner, while amplifying feature activation with greater intensity. Compared with several other pooling approaches, it is differentiable and may keep more data in the activation map after downsampling. The following equation illustrates how soft pool is calculated:

$$\tilde{a}_i = \sum_{i\in R}\frac{e^{a_i}a_i}{\sum_{j\in R}e^{a_j}} \tag{10}$$

where R is the 2-D space area whose size is equal to the size of the pool specific kernel, and a_i is the pixel in R.

Subsequently, the features after soft pool are passed through a 1×1 convolutional layer increasing the dimension to output $F_2 \in \mathbb{R}^{h/2\times w/2\times 2c_1}$. And this is a representation of the F_2

$$F_2 = \phi(\text{SoftPool}(a)) \tag{11}$$

where $\phi(\cdot)$ is a 1×1 convolution layer with batch normalization and GELU.

These two branches not only collect small-scale objects, but also obtain finer details, which are also very important in downsampling. Therefore, the superposition operation of these two branches as the output S of FDM can be expressed as follows:

$$S = F_1 \bigoplus F_2 \tag{12}$$

where $\bigoplus$ stands for element-level addition.

F. Projection and Concatenation

In our approach, to equalize the number of channels, the output of each temporal transformer at each stage is fed into an MLP module. Its output is then upsampled to the same size of $\frac{H}{16} \times \frac{W}{16}$, and connected to the fusion modules from different stages, the connected feature map is $\frac{H}{16} \times \frac{W}{16} \times 3C$. Finally, the result of the concatenation is generated by convolution operation with size of $\frac{H}{16} \times \frac{W}{16} \times C$ and then fed into the final transformer encoder block for encoding. The above operation can be formulated as follows:

$$\hat{Z}_i = \text{MLP}_{C_i\to C}(F_i), i = 1, 2, 3 \tag{13}$$

$$\hat{F}_i = \text{Upsample}_{\frac{H}{16}\times\frac{W}{16}}(\hat{F}_i), i = 1, 2, 3 \tag{14}$$

$$F_i = \text{Conv}_{3C\to C}[\hat{F}_1; \hat{F}_2; \hat{F}_3]. \tag{15}$$

G. Fusion Module

As shown in Fig. 1, when extracting bi-temporal multiscale features, we use the fusion module to treat the two images before and after the change as a whole. The fusion module in the model is as follows:

$$F_{\text{fus}} = \text{BN}(\text{RELU}(\text{Conv}_{3\times3}[F_i^{\text{pre}}; F_i^{\text{post}}])). \tag{16}$$

Obviously, the fusion module consists of 3×3 convolution operation, ReLU and BatchNorm2d (BN), where F_i^{pre} and F_i^{post} are the feature representations of multilevel features of prechange and postchange images. The fusion method adopted by some previous methods is to calculate the absolute difference among them, such as Transcd [54] and BIT [55]. However, the decoder undergoes several downsampling operations to obtain different scale features, some spatial and detail information is lost. We introduce the fusion module generated by the encoder into the decoder, thus mitigating the loss of information, and adopt a method of adding channel fusion in the merger block to facilitate exploring the differences and potential associations between bitemporal images.

IV. Experiments

A. Datasets

We use three different CD datasets to demonstrate the effectiveness of our method, they are LEVIR-CD [48], DSIFN-CD [59], and CDD [56] datasets.

The LEVIR-CD [48] is RS images with 1024×1024 resolution, which is an open, comprehensive building CD dataset. It focuses on changes that affect buildings, such as building expansion and building decline. We crop the images to a size of 256×256, but not overlapping each other, we obtain 7120/1024/2048 samples for train/val/test for these images, respectively.

A broad RS CD dataset that contains the alterations to various land-cover items is the DSIFN-CD [59] dataset. It comprises of six sizable, high-resolution, bitemporal images that each represent one of six Chinese cities. We crop 512×512 images to size 256×256, this process has no overlapping regions, resulting in 14400/1360/192 data for train/val/test.

There are 11 pairs of multispectral photos in the CDD [60] dataset, including four pairs of 1900×1000 pixel images and seven pairs of 4725×2200 pixel images from various seasons. These changes in the CDD dataset are primarily brought on by structures, such as buildings, car, and roads. All images are rotated and cropped into 256×256 image patches. For these patches, 10 000/3000/3000 samples for train/val/test, respectively.

B. Implementation Details

We use NVIDIA RTX 2080Ti for training and PyTorch to implement our model. We use random flip, random rescale (0.8–1.2), random crop, Gaussian blur, and random color jittering to enhance the data. We train the model with weight decay of 0.01 and beta values of (0.9, 0.999) using the cross-entropy (CE) loss and AdamW optimizer. The learning rate starts out at 0.0001 and linearly declines to 0 after 200 training epochs. The model is trained with a batch size of 8.

C. Evaluation Criteria

The scores of precision (OA), intersection of union (IoU), precision (P), recall (R), and F1-score (F1) are used in our experiments to evaluate the performance of the model. These are all common metrics in binary classification, and F1 is a comprehensive evaluation, which measures the average of P and R. IoU is expressed as the overlap of the pixel areas of the predicted results and true map. Therefore, these two indicators are generally used to evaluate the experimental effect in CD. The following is the IoU calculating formula:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{17}$$

Here, is how the F1 score is calculated

$$F1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \tag{18}$$

where P = TP/(TP + FP) and recall = TP/(TP + FN). TP, TN, FP, and FN represent the number of true positive, true negative, false positive and false negative, respectively.

D. Comparison to State-of-the-Art

We compare GateFormer with several excellent methods, they are FC-EF [23], FC-Siam-conc [23], FC-Siam-diff [23], STANet [48], DTCDSCN [26], BIT [55], ChangeFormer [56], FTN [51], ICIF-Net [57], and DMI-Net [61].

1) *FC-EF* [23]*:* Bitemporal original images are concatenated in the initial stage and extract features through the U-net structure.
2) *FC-Siam-conc* [23]*:* The decoder generates bitemporal features at various levels by extracting features through the Siamese convnet. Finally, the bitemporal information is fused by feature cascade and decoder of it leverages obtained hierarchical information to gradually generate change information.
3) *FC-Siam-diff* [23]*:* FC-Siam-diff is a U-shaped structure, just like FC-Siam-conc. The distinction is that FC-Siam-diff merges the bi-temporal characteristics using absolute difference approach.
4) *STANet* [48]*:* The CD self-attention module is added when Resnet18 is used to extract features, which calculates the attention weights of any two pixels at different times and positions, and uses them to generate more distinctive features.
5) *DTCDSCN* [26]*:* It is based on dual attention, which makes use of the interdependence of channel and spatial information to extract features. In addition to the CD task, two additional semantic segmentation decoders are trained to obtain object-level data.

TABLE I
EXPERIMENTAL RESULTS ON THREE DIFFERENT TEST SETS

method	LEVIR-CD					DSIFN-CD					CDD				
	Precision	Recall	F1	IoU	OA	Precision	Recall	F1	IoU	OA	Precision	Recall	F1	IoU	OA
FC-EF [23]	86.92	80.15	83.46	71.51	98.33	72.60	52.74	61.13	43.98	88.62	86.06	54.23	66.53	49.85	93.56
FC-Siam-conc [23]	91.99	76.74	83.72	71.98	98.53	66.51	54.41	59.73	42.56	87.60	86.87	71.20	78.38	64.45	95.35
FC-Siam-diff [23]	89.58	83.28	86.31	75.94	98.67	59.67	65.71	62.54	45.50	86.63	90.27	56.90	69.80	53.61	94.19
STANet [48]	83.81	**91.00**	87.26	77.40	98.66	67.71	61.68	64.56	47.66	88.49	93.10	93.90	93.50	87.82	98.30
DTCDSCN [26]	85.53	86.83	87.63	78.05	98.77	53.87	77.99	63.72	46.76	84.91	95.12	92.44	93.76	88.25	98.55
BIT [55]	89.24	89.37	89.31	80.68	98.92	68.36	70.18	69.26	52.97	89.41	94.85	94.10	94.48	89.53	98.70
ChangeFormer [56]	92.11	88.67	90.36	82.41	99.04	89.10	86.05	87.55	77.86	95.84	94.17	93.78	93.97	88.63	98.58
FTN [51]	92.53	89.01	90.65	82.91	99.05	**93.44**	89.45	**91.23**	82.20	**97.18**	89.70	87.26	88.46	79.31	97.17
ICIF-Net [57]	90.70	89.50	90.09	81.97	98.99	84.32	85.53	84.92	73.80	94.80	95.21	94.62	94.91	90.30	98.80
DMI-Net [61]	**92.78**	88.89	90.59	82.94	99.06	81.27	91.37	89.27	80.63	96.27	**94.98**	94.93	95.00	90.43	98.81
GateFormer	91.16	90.28	**90.72**	**83.02**	**99.06**	91.63	**89.50**	90.75	**82.73**	96.83	95.01	**95.04**	**95.03**	**90.52**	**98.83**

All the scores are described in percentage (%).
The bold entities represents the best performing data.

6) *BIT* [55]*:* It leverages the transformer to connect semantic concepts in token-based time and space. To obtain the change map, BIT learns a small number of tokens to express high-level concepts that reflect the change of interest contained in the bi-temporal image.
7) *ChangeFormer* [56]*:* It is a transformer-based Siamese network structure that combines a MLP decoder with a hierarchically structured transformer encoder in order to provide correct change information.
8) *FTN* [51]*:* It is the architecture of a FTN, which aggregates multiple visual features in a pyramid fashion to improve feature representation.
9) *ICIF-Net* [57]*:* It is an intrascale cross interaction and interscale feature fusion network, and uses the local and global information extracted by CNN and transformer, in order to thoroughly aggregates local and global features.
10) *DMI-Net* [61]*:* It is an encoder–decoder-based architecture, which mainly proposes intertemporal joint-attention block to guide the global feature distribution of each input, and then the differential features are extracted using subtraction and concatenation.

The comparison results for the test sets LEVIR-CD, DSIFN-CD, and CDD are given in Table I. According to the quantitative findings, our GateFormer model consistently surpasses different approaches across various datasets by a substantial margin. In particular, our GateFormer improves recent ChangeFormer in F1/IoU by 0.2%/0.3%, 2.4%/3.8%, and 0.8%/1.4% for LEVIR-CD, DSIFN-CD, and CDD, respectively. At the same time, we can see from the table that the recent models using transformer as the backbone, such as BIT, ChangeFormer, FTN, ICIF-Net, and our GateFormer can achieve outstanding performance than CNN. It might be related to the transformer's capacity for modeling the context. Meanwhile, this confirms that CNN-based models have some limitations in global feature extraction as well as in establishing long-range information dependencies.

We summarize extensive experiments on three bitemporal RS image CD datasets. The overall performance of LEVIR-CD, DSIFN-CD, and CDD test samples is reported in Table I, where the quantitative results show that our GateFormer outperforms other competitors by a significant margin in terms of the five metrics as a whole. To see it more clearly, we mark the different regions with red dotted boxes in the visualizations of the three datasets. And we can also see that the GateFormer model outperforms others.

1) Visualization on LEVIR-CD: For the LEVIR dataset, buildings are the main cause of changes, and the detail processing of building edges is also more advantageous for our GateFormer, as shown in Fig. 5(a) and (d). As our method uses the gate mechanism, it pays more attention to details and global features, and has a clearer edge shape in complex buildings. For the strong seasonal and illumination changes in Fig. 5(b) and (c). FC-EF, FC-Siam-Di, and FC-Siam-conc have missed detection phenomena. Compared with the traditional method using convolution, the method using transformer can find subtle changes more obviously. GateFormer is more able to identify accurate building changes. Fewer false and miss detections validate the ability of our GateFormer to detect changes in scenarios with large variations.

2) Visualization on DSIFN-CD: The DSIFN-CD dataset is also a building-dominated dataset. However, compared with the LEVIR-CD dataset, it has a larger variance in architectural features. For the small changes existing in Fig. 6(a), some other methods directly ignore the existence of buildings, and even other methods mistakenly treat the road as the change area from the visualization results, but GateFormer could still detect it. At same time, in Fig. 6(a) and (d), some of the compared methods cannot fully identify the huge building area due to the limited receptive field. However, our GateFormer model provides more thorough results. Last, in Fig. 6(b) and (c), it is easy to see the result of detection that GateFormer can recover the details of the

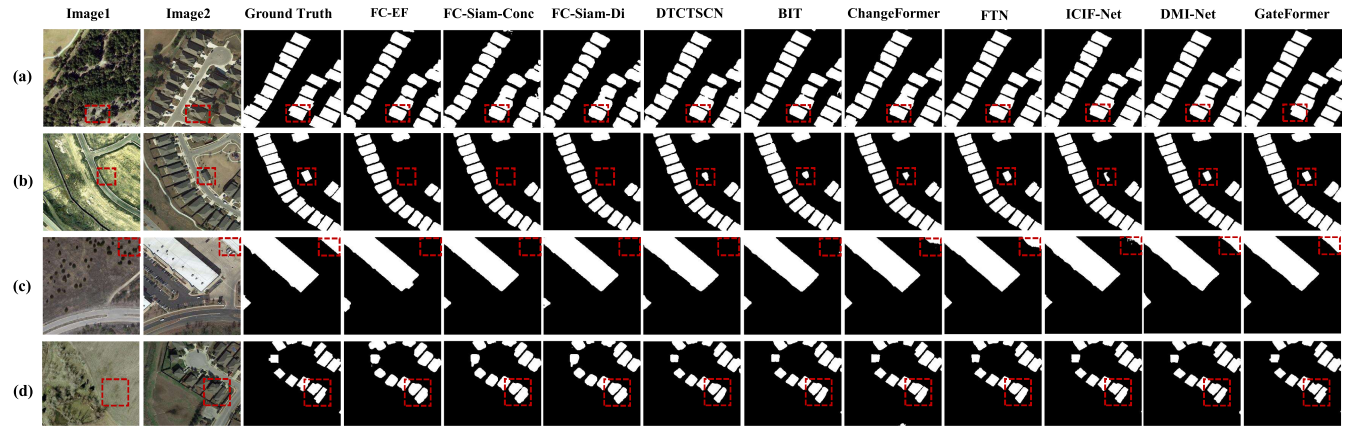


Fig. 5. Visualization results of different methods on the LEVIR-CD test set. All of the comparable prediction results for various samples are shown in (a)–(d), respectively. We have highlighted the key areas with red dotted boxes.

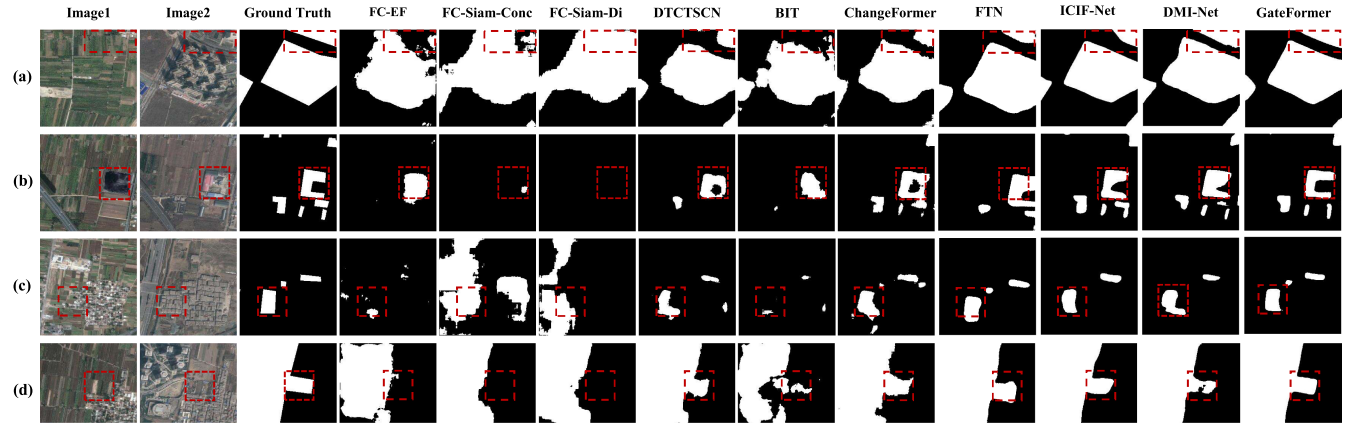


Fig. 6. Visualization results of different methods on the DSIFN-CD test set. All of the comparable prediction results for various samples are shown in (a)–(d), respectively. We have highlighted the key areas with red dotted boxes.

changed area and keep the edge of the changed area, which is closer to the ground truth, and our model can more effectively detect the pseudovariation.

3) Visualization on CDD: The CDD dataset includes a variety of constantly changing objects. The changes brought about by cars and buildings are shown in Fig. 7(a). Fig. 7(b) and (c) show changes caused by roadways, and the changes in Fig. 7(d) are due to different buildings. As you can see that GateFormer can predict changes more accurately than other methods. For example, in Fig. 7(b) and (c), road changes detected by GateFormer are more coherent and complete. In Fig. 7(d), the boundary between road and house changes is more obvious. This also further indicates that the GateFormer method has stronger semantic discrimination ability.

To compare the effectiveness of the models, Table III gives the number of model parameters (Params.), the floating-point operations per second (FLOPs), and the size of the image used is $256 \times 256 \times 3$. Among them, the network model's complexity decreases with decreasing FLOPs and Params. Due to parallel feature extraction in transformer, GateFormer has strong global modeling ability and semantic representation ability, but also sacrifices a lot of resource requirements. In the future, we will continue to study and explore the development of lightweight models.

TABLE II
ABLATION EXPERIMENTS OF THREE IMPORTANT MODULES ON THE DSIFN-CD DATASET

Method Name	MIoU	Ave.F1
TransU	76.20	86.42
TransU+GAM	77.03	87.02
TransU+SPM	78.56	88.22
TransU+FDM	80.73	89.37
TransU+SPM+GAM	79.23	88.53
TransU+SPM+FDM	80.94	89.55
TransU+GAM+FDM	81.87	90.23
TransU+SPM+GAM+FDM	82.73	90.75

E. Ablation Study

Throughout the experiment, the transformer is configured as follows: the patch size is 7, the number of layers corresponding

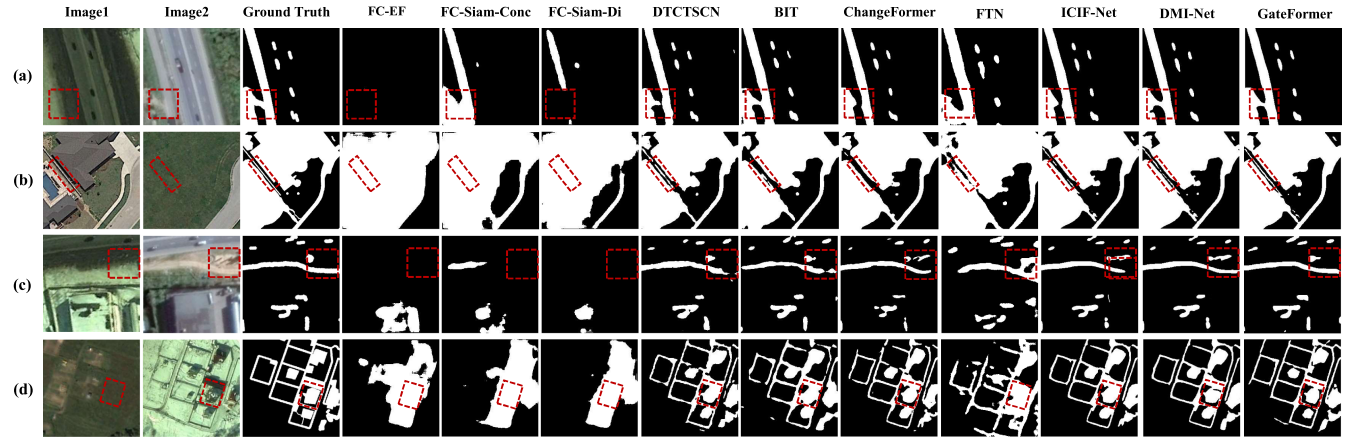


Fig. 7. Visualization results of different methods on the CDD test set. All of the comparable prediction results for various samples are shown in (a)–(d), respectively. We have highlighted the key areas with red dotted boxes.

TABLE III
MODEL EFFICIENCY COMPARISON

Method Name	Params. (M)	FLOPs (G)
FC-EF [23]	10.35	3.57
FC-Siam-conc [23]	10.55	5.32
FC-Siam-diff [23]	10.35	4.72
STANet [48]	16.93	6.58
DTCDSCN [26]	42.26	13.21
BIT [55]	3.55	10.35
ChangeFormer [56]	117	21.18
FTN [51]	677.9	45.23
ICIF-Net [57]	52.83	25.36
DMI-Net [61]	48.22	53.22
GateFormer	623.2	32.17

We list the number of parameteres (Params.) and FLOPs. calculate these with a size of 256 × 256 × 3 image size.

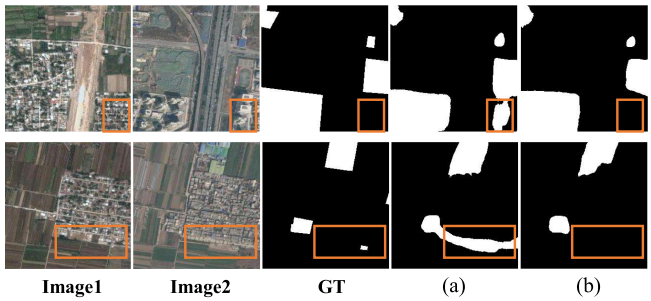


Fig. 8. Comparison of visualization results before and after applying GAM in TransU framework. (a) TransU. (a) TransU. (b) TransU + GAM.

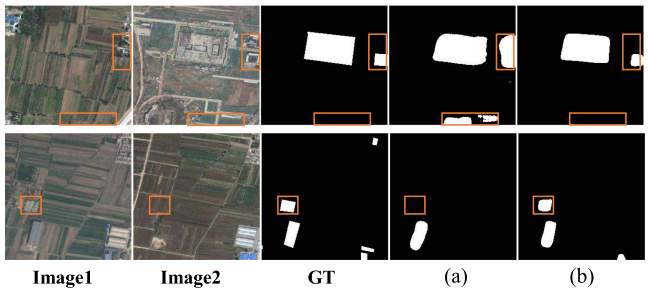


Fig. 9. Comparison of visualization results before and after applying SPM in TransU framework. (a) TransU. (b) TransU + SPM.

to each stage is {3, 4, 6, and 3}, and the number of heads corresponding to each layer is {1, 2, 4, and 8}. In order to prevent the fluctuations generated by the ablation experiment and ensure its accuracy, we conduct five experiments and select average values for IoU and F1, respectively. The evaluation metrics are MIoU and Ave.F1. We perform ablation experiments using TranU as a baseline to assess the effectiveness of the proposed network structure and three crucial modules. TransU is a model of GateFormer structure that removes GAM, FDM, and SPM modules, which uses transformer for feature extraction.

1) Effect of GAM: Table II demonstrates that when the GAM is taken into account within the TransU framework, the segmentation results rise by 0.83% on MIoU and 0.60% on Ave.F1. We can see the visualization results intuitively in Fig. 8. In the first row, the errors caused by buildings with different appearances are avoided after using GAM. Here, we identify the building as unchanged before and after the change, although the appearance of the building has changed. And in the second row, vegetation in different seasons may also lead to wrong prediction, but it can be identified after adding GAM. It is shown that greater global context information and semantic recognition abilities occurred after employing GAM.

2) Effect of SPM: In Table II, as a result of applying SPM to the TransU framework, MIoU grows by 2.36%, and Ave.F1 rises by 1.80%. As shown in the first row of Fig. 9, the module mistake roads as changes and the boundary recognition of small objects is not clear enough. In the second row, the houses close to the ground color and "house" is embedded in "farmland," which makes the model unable to identify. The visualization results in Fig. 9(b) show that implementing of SPM successfully prevents judgment errors involving fuzzy semantics.

TABLE V
ABLATION EXPERIMENTS OF THE DIFFERENT SIZE OF THE DILATION RATE IN SPM ON THE DSIFN-CD DATASET

Dilation Rate	IoU	F1
-	82.24	90.21
[1, 2]	82.60	90.52
[1, 3]	82.48	90.40
[1, 4]	82.70	90.72
[2, 3]	82.57	90.45
[3, 4]	82.55	90.67
[2, 4]	82.73	90.75

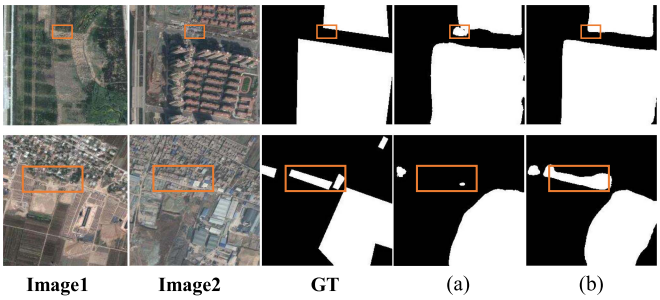


Fig. 10. Comparison of visualization results before and after applying FDM in TransU framework. (a) TransU. (a) TransU. (b) TransU + FDM.

In Table V, we also verify the influence of different dilation rate in SPM's dilated convolution on the experimental results. To account for the decreasing resolution and the number of parameters during downsampling. We set the dilation rate of any two experimental comparison of 1 to 4. In Table V, we can see that when the dilation rate of [1,4] parallel combination on the DSIFN-CD dataset is the highest accuracy. In addition, we also set up an experiment without dilated convolution. It is obvious that when SPM is used to extract features, after applying dilated convolution, the receptive field is expanded, resulting in the collection of more deep feature information and a markedly enhanced experimental result.

3) Effect of FDM: Table II demonstrates that when the FDM is used independently, the model improves MIoU and Ave.F1 by 4.53% and 2.95%, respectively. These results of the visual comparison are shown in Fig. 10. For the processing of edge information in the first row and the extraction of small and dense objects in the second row, FDM can perfectly outperform the TransU model. According to this result, FDM tends to increase the detection precision of small-scale ground objects and edge information.

In addition, we also study the effect of cooperation between different module under the TransU-based framework. As given in Table II, when SPM and GAM are introduced at the same time, MIoU and Ave.F1 increase by 3.03% and 2.11%, respectively. The detection result is increased by 4.76% MIoU and 3.31% Ave.F1 when SPM and FDM are both used. When GAM and FDM are taken into account, the increases for MIoU and Ave.F1 are 5.67% and 3.81%, respectively. In comparison with TransU, our GateFormer's three crucial modules (SPM, GAM, and FDM) result in increase of 6.53% on MIoU and 4.33% on Ave.F1. From this, we can also see the positive effect of the three modules of GateFormer.

TABLE IV
ABLATION EXPERIMENT OF THE SOFTPOOL BRANCH ON THREE DATASETS

Softpool	LEVIR-CD		DSIFN-CD		CDD	
	F1	IoU	F1	IoU	F1	IoU
✓	90.70	82.98	90.75	82.73	95.03	90.52
×	90.31	82.45	89.81	81.69	94.30	89.20

4) Effect of softpool branch: To assess the performance of the softpool branch, we compare and verify with and without softpool branch on three experimental datasets. We can find that F1 and IoU can obtain the best performance on the three datasets, respectively, after adding softpool branch. Obviously, as given in Table IV, compared with not adding softpool branch. After effective use of softpool, the F1/IoU of LEVIR-CD, DSIFN-CD, and CDD can be improved by 0.39%/0.53%, 0.94%/1.04%, and 0.73%/1.32%, respectively. The experiment also proves that we can get more refined features after using soft-pool. As a result, misjudgments and spurious changes are reduced.

V. CONCLUSION

In this article, we propose a combination of transformer and Siamese U-shaped with GAM model named GateFormer. It aims to obtain the comprehensive context information and detailed features of RS images. More specifically, the GateFormer employs a GAM to connect the encoder and decoder before skipping connections. It filters the low-level information by guiding the high-level information, making it easier to capture the region of interest. In addition, SPM and FDM are further intended to enhance transformer's capacity for global modeling. The SPM alleviates the recognition error caused by occlusion by aggregating 2-D feature information. The FDM also saves as much detail as possible during downsampling. In the experiments, we use three RS CD public datasets, they are LEVIR-CD, DSIFN-CD, and the CDD datasets. From these experimental results, it is clear that GateFormer is significantly better than other methods.

REFERENCES

[1] S. Yin et al., "A review of the research progress of multi temporal remote sensing image change detection methods," *Spectrosc. Spectral Anal..*, vol. 33, pp. 3339–3342, 2013.
[2] 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.
[3] J. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, "Building damage detection in satellite imagery using convolutional neural networks," 2019, *arXiv:1910.06444*.
[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] N. Coops, M. Wulder, and J. White, "Identifying and describing forest disturbance and spatial pattern: Data selection issues and methodological implications," *Forest Disturbance Spatial Pattern: Remote Sens. GIS Approaches*, vol. 2, no. 264, pp. 31–61, Jul. 2007.

[7] P. Howarth and G. Wickware, "Procedures for change detection using landsat digital data," *Int. J. Remote Sens.*, vol. 2, no. 3, pp. 277–291, 1981.

[8] A. Ludeke, R. Maggio, and L. Reid, "An analysis of anthropogenic deforestation using logistic regression and gis," *J. Environ. Manage.*, vol. 31, no. 3, pp. 247–259, 1990.

[9] Y. Bayarjargal, A. Karnieli, M. Bayasgalan, S. Khudulmur, C. Gandush, and C. Tucker, "A comparative study of NOAA–AVHRR derived drought indices using change vector analysis," *Remote Sens. Environ.*, vol. 105, no. 1, pp. 9–22, 2006.

[10] J. Deng, K. Wang, Y. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, 2008.

[11] G. Zhang, G. Li, and W. Cui, "High-resolution remote sensing change detection by statistical-object-based method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2440–2447, Jul. 2018.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time 0bject detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2016, pp. 779–788.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

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

[16] 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.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[18] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell..*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[19] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput.-Assist. Interv.*, 2015, pp. 234–241.

[21] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[22] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett..*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

[23] R. Daudt, B. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[24] J. Liu, W. Xuan, Y. Gan, Y. Zhan, J. Liu, and B. Du., "An end-to-end supervised domain adaptation framework for cross-domain change detection," *Pattern Recognit.*, vol. 132, 2022, Art. no. 108960.

[25] L. Song, M. Xia, J. Jin, M. Qian, and Y. Zhang, "Suacdnet: Attentional change detection network based on siamese u-shaped structure," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, 2021, Art. no. 102597.

[26] 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.

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

[28] 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.

[29] W. Gao, Y. Sun, X. Han, Y. Zhang, L. Zhang, and Y. Hu, "AMIO-Net: An attention-based multiscale input–output network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, no. 5, May 2023.

[30] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

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

[32] 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.

[33] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[34] A. Dosovitskiy et al., "An image is worth 16x16 words:transformers for image recognition at scale," *Proc. Inf. Conf. Adv. Neural Inf. Process. Syst.*, vol. 12, no. 10, pp. 1–8, Jan. 2020.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[36] C. Chen, Q. Fan, and R. Panda, and S. Zagoruyko, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2021, pp. 357–366.

[37] J. Gu and S. Zagoruyko, "Multi-scale high-resolution vision transformer for semanti segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12084–12093.

[38] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 17864–17875.

[39] Z. You, J. Wang, S. Chen, J. Tang, and B. Luo, "FMWDCT: Foreground mixup into weighted dual-network cross training for semisupervised remote sensing road extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5570–5579, Jun. 2022.

[40] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8739–8748.

[41] K. Lin et al., "SwinBERT: End-to-end transformers with sparse attention for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17928–17937.

[42] Q. Hou, L. Zhang, M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4003–4012.

[43] 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.

[44] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.

[45] 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, Jun. 2019, Art. no. 1382.

[46] 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, no. 8, pp. 1194–1206, Aug. 2021.

[47] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," in *IEEE Geosci. Remote Sens. Lett.*, vol. 19, no. 10, pp. 1–5, Aug. 2021.

[48] 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, May 2020, Art. no. 1662.

[49] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5604816.

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

[51] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proc. Asian Conf. Comput. Vis.*, 2023, pp. 75–92.

[52] 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.

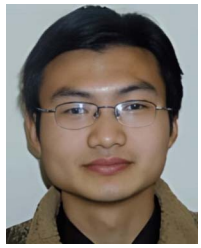
[53] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 228–239, 2022.
[54] Z. Wang, Y. Zhang, L. Luo, and N. Wang, "Transcd: Scene change detection via transformer-based architecture," *Opt. Exp.*, vol. 29, no. 25, pp. 41409–41427, 2021.
[55] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5607514.
[56] W. Bandara and V. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2022, pp. 207–210.
[57] 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, no. 10, pp. 1–5, Oct. 2022.
[58] A. Stergiou, R. Poppe, and G. Kalliatakis, "Refining activation downsampling with softpool," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10337–10346.
[59] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.
[60] M. Lebedev, Y. Vizilter, O. Vygolov, V. Knyaz, and A. Rubis, "Change detection in remote sensing images using conditional adversarial networks. international archives of the photogrammetry," *Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, 2018.
[61] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.

Li-Li Li received the B.S degree in computer science from the Changchun Institute of Technology, ChangChun, China, in 2020. She is currently working toward the master's degree in computer technique from Anhui University, Hefei, China.

Her research focuses on remote sensing change detection.

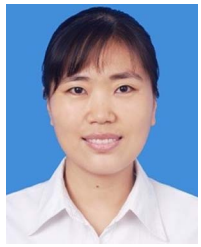
Zhi-Hui You received the B.S degree in computer science from Jiangxi Agricultural University, NanChang, China, in 2020. He is currently working toward the Ph.D. degree in computer science from Anhui University, Hefei, China.

His research interests include machine learning, pattern recognition, semantic segmentation, and change detection in remote sensing.



Si-Bao Chen (Member, IEEE) received the B.S. and M.S. degrees in probability and statistics, and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 2000, 2003, and 2006, respectively.

From 2006 to 2008, he was a Postdoctoral Researcher with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei. Since 2008, he has been a Teacher with Anhui University. From 2014 to 2015, he was a Visiting Scholar with The University of Texas at Arlington, Arlington, TX, USA. His research interests include image processing, pattern recognition, machine learning, and computer vision.



Li-Li Huang received the B.E. and master's degrees from Anhui University, Hefei, China, in 2010 and 2013, respectively, and the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2019, all in computer science.

She is currently a Teacher with the School of Computer Science and Technology, Anhui University. Her current research interests include artificial intelligence, computer vision, and cognitive science.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively.

He is currently a Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, machine learning, and deep learning.

Bin Luo (Senior Member, IEEE) received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, Heslington, U.K., in 2002.

From 2000 to 2004, he was a Research Associate with the University of York. He is currently a Professor with Anhui University. His research interests include graph spectral analysis, large image database retrieval, image and graph matching, statistical pattern recognition, digital watermarking, and information security.