IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 July 2023, accepted 18 August 2023, date of publication 22 August 2023, date of current version 26 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307642

RESEARCH ARTICLE

DAHT-Net: Deformable Attention-Guided Hierarchical Transformer Network Based on Remote Sensing Image Change Detection

GANG SHI, YUNFEI MEI, XIAOLI WANG, AND QINGWEN YANG
School of Information Science and Engineering, Xinjiang University, Ürümqi 830017, China

Corresponding author: Gang Shi (shigang@xju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62162059, and in part by the Third Xinjiang Scientific Expedition Program under Grant 2021xjkk1400.

ABSTRACT Remote sensing image change detection (CD) refers to the automated or semi-automated detection of differences between two remote sensing images taken at different times in the same region. To achieve better global modeling and faster inference, we propose a network architecture containing a hierarchical swin transformer block and deformable attention transformers crossed for encoding and lightweight MLP decoding to solve the CD task. The deformable attention transformer allows adaptive adjustment of the relationships and weights between feature mappings to effectively combat variations and noise interference in various scenes. The alternating use of swin transformer block and deformable attention transformer ensures the efficiency as well as the flexibility of the model. The lightweight MLP approach provides better ability to extract spatial features and contextual information, as well as faster inference speed. Compared with other methods, our proposed DAHT-Net method improves F1 scores by 0.98 and 2.61 on LEVIR-CD, CDD and two publicly available benchmark datasets, respectively, and performs well on other measures. These experimental results validate that the DAHT-Net network outperforms other comparative methods and highlight its effectiveness in remote sensing image change detection. In summary, our proposed hierarchical deformable attention-guided transformer network model provides a promising solution for remote sensing image change detection with superior performance compared to other state-of-the-art methods.

INDEX TERMS Change detection, global modeling, hierarchical transformer, deformable attention.

I. INTRODUCTION

In order to quantitatively analyze and determine the features and processes of land surface changes, it is necessary to use multi-temporal remote sensing data and employ image change detection(CD [1]) methods to extract change information. These methods have been widely used in a range of fields, including urban planning and layout, land cover and change monitoring, as well as dynamic target monitoring in military reconnaissance, such as roads and bridges. Many previous CD approaches have utilized Convolutional Neural Networks (CNN) due to their powerful feature representation capabilities. However, such techniques primarily employ attention re-weighting in channel and spatial dimensions to obtain dual temporal features, which result in the loss of valuable information in the Pooling layer while ignoring the correlation between the local and global information. Currently Generative Adversarial Network(GAN) models have achieved some success in image change detection and are able to perform unsupervised learning to deal with the temporality of remote sensing images from different sensors, e.g. [2] and [3]. However, research on GAN networks for image change detection is still in its infancy.

Nowadays, Transformer [4] model has become a new paradigm in the field of natural language processing (NLP), and more and more researchers try to apply the powerful modeling ability of the Transformer model to the field of computer vision (CV). The Transformer's self-attention mechanism, with its advantages of domain independence and efficient

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang.

computation and learning of long-distance dependencies, is introduced to image and related cross-modal domains.

Intuitively, the aforementioned advantages of Transformer are well suited to overcome some of the drawbacks of CNN-based approaches. Nowadays, more and more researchers are improving Transformer's network structure to CV tasks with outstanding results, e.g., [5], [6], [7], [8], and [9]. Nevertheless, the attention is computed as a squared term of the input sequence length, so the computational cost and memory required for inputting higher-resolution images are very large. In the image domain, variable convolution is a powerful and effective method for attention to sparse spatial locations, which can effectively avoid the above problems. If variable convolution is used exclusively it can lead to a lack of modeling mechanism between elements. Therefore, there is a need to rely on the sparse attention of the data to flexibly model the relevant features, leading to the deformable mechanism first proposed in DCN [10].

To address these issues, we designed a transformer base module that uses different attention at different levels to make CD task processing efficient and accurate and to exchange semantic information between features. The first two levels use shift-window attention in Swin Transformer to enhance the information exchange between features, and the last two levels use variable attention to enhance the inter-element modeling capability.

The main contributions of this paper are as follows:

- We propose a hierarchical Transformer network model, named DAHT-Net, as the backbone of our approach. This model optimizes computational efficiency and minimizes memory usage by reducing the number of operations required to calculate the association between distant locations. Through a hierarchical approach, we are able to extract more global information, improve detection accuracy, and demonstrate strong robustness to pseudo-change cases.
- In the DAHT-Net network architecture, we propose to use swin transformer block (STB) for the first two layers and deformable attention transformer (DAT) as encoder for the last two layers. This approach allows a more adaptive and accurate modeling of the features within the network.
- In the DAHT-Net network structure, we propose to use a lightweight MLP (LMLP) as a decoder for fast classification and improved recognition rate to predict change maps.

We will first present related work in Section II. Then in Section III the methodology used for our DAHT-Net network model is presented in less detail. In Section IV, we present our experimental results as well as the ablation experiments. Finally, in Section V, we summarize our work and what we would like to investigate in the future.

II. RELATED WORK

In this section, a survey is presented on the application of CNN, Transformer, and DCN [10] in image change detection. Their strengths and weaknesses in this field are summarized and a comparative analysis is provided.

A. CNN IN IMAGE CHANGE DETECTION

Many previous CD methods are based on CNN due to their strong feature representation capability. CNN-based CD methods usually enhance the semantic representation capability of the network by changing the network structure, optimizing the loss function, and adding an attention mechanism. In terms of network structure, Zhan et al. [11] was the first to use Siamese convolutional networks that can process dual-time images in parallel to handle CD tasks. Subsequently, a large number of CD methods using Siamese convolutional network structures have been proposed.

Daudt et al. [12] designed the first end-to-end training CD method by proposing three models, namely FC-EF [12] based directly on U-Net [13], FC-Siam-conc [12] and FC-Siam-diff [12] with two twin neural network structures based on FC-EF [12]. FC- EF [12] implements the early fusion method by connecting differential mappings of dual-temporal features in the decoding phase. The FC-Siam-conc [12] and FC-Siam-diff [12], on the other hand, directly concatenate the dual-temporal features to achieve the later fusion method.

However, these methods still have difficulty in extracting global information in space-time due to the inherently local nature of convolutional operations. The most intuitive way to reduce the inherent localization of convolutional operations is to increase the receiver field. Therefore, Zhang et al. [14] used dilated convolution instead of traditional convolution and achieved some results. Chen and Shi [15] proposed STANet and Chen et al. [16] proposed DASNet, and they used Resnet18 and Resnet50 as backbone networks, respectively. Compared with shallow networks, STANet and DASNet have stronger feature extraction capabilities.

It is well known that dense connectivity between features [17] can improve network performance. In [18] and [19], the authors added dense connections between features of different layers to enhance the capabilities of CD networks. Although attention-based methods are effective in capturing global details, they have difficulty in linking remote details spatiotemporally because they use attention to reweight the dual temporal features obtained through convolutional networks in both channel and spatial dimensions.

B. TRANSFORMER IN IMAGE CHANGE DETECTION

To enhance the Transformer neural network's representation capabilities in computer vision tasks, researchers proposed adding a fully graph-based attention mechanism that considers global information. Originally developed for natural language processing, the Transformer neural network employs a self-attentive mechanism with remarkable feature representation capabilities. It has achieved comparable, if not superior, performance to traditional convolutional neural networks (CNNs) [20] in various computer vision tasks,

TABLE 1. Pros and cons of all related work.

related work	pros	cons
CNN in image change detection [11]	Local feature learning; Parameter sharing; Non-linear modeling capability; Data enhancement	Sensitivity to small-scale variation; Large training data requirement; Poor interpretability; Higher computing resource requirements
Transformer in image change detection [21]	Global relationship modelling; Multilevel feature representation; Parallel computing	Large training data requirement; Higher computing resource requirements; Poor interpretability; Challenges of handling large images
DCN in Transformer [10]	More sensitive to deformation and scale changes; Better position perception; Enhanced modeling capability	Increase in computational complexity; Parameter addition; More training data needed

such as image classification (ViT [21]), detection (DETR [7]), and segmentation (SETR [22]).

Notably, unlike prior approaches that integrate attention mechanisms with CNNs [20] or replace specific CNN components [20], the Visual Transformer (ViT [21]) represents the first pure Transformer approach for image classification tasks. This approach has demonstrated excellent performance results and has shown scalability.ViT [21] takes as input 2D image patches with positional embedding and pre-trains them on large datasets without relying on convolution. On the other hand, DETR [7] uses transformers as encoders and decoders, greatly simplifying the framework for target detection. Transformer networks have a larger effective receptive domain, providing more powerful context modeling capabilities than convolutional neural networks between any pair of pixels in an image. The more popular Transformers have recently shown their powerful performance in CV (image classification, segmentation), such as ViT [21], SETR [22], Swin [23], Twins [8], and SegFormer [9]. Although the Transformer has a larger receptive domain and greater context-shaping capability, little work has been done on CD.

The Transformer structure is similar to the ConvNet encoder (ResNet18) is used in combination to enhance the feature representation while maintaining the overall ConvNet-based feature extraction process.

C. DCN IN TRANSFORMER

Compared with the global and dense attention mechanism of Transformer, the improvement of DCN [10] can use each reference point to focus only on a set of sampling points in the neighborhood, thus achieving a local and sparse efficient attention mechanism. Deformable convolutions can learn sparse spatial locations, but they also lack relational modeling capabilities, which happens to be what Transformer does best, so Zhu et al. [24], Zhu et al. [25], Xia et al. [5] proposed to apply DCN to a transformer and achieved good results.

III. METHODOLOGY OF DAHT-NET

In this section, the methodology of our proposed approach is presented. Firstly, the general architecture of our network is introduced, which consists of an encoder module called Swin Transformer Block (STB), a deformable attention module called Deformable Attention Transformer (DAT), and a decoder module called Lightweight Multi-Layer Perceptron (LMLP). Next, the details of each component and their functions are elaborated upon. Lastly, the loss function used in our experiments is introduced.

A. OVERALL

As with most binary CD methods, the network input is a pair of aligned dual-temporal images, denoted as T1 and T2, with dimensions H $\times$ W $\times$ C, where C is the number of channels, resulting in a change map with a number of channels of 1 and the same height and width as the input image. For each pixel in the change map, 1 means that a change occurs and 0 means that no change occurs. In this paper, a hierarchical transformer with variable attention that constitutes a twin network is proposed for extracting the global information of the dual-time image to handle the CD task.

The overall architecture of this network is shown in Figure 1, and the encoder-decoder structure is used for the whole network architecture. In addition, a variable attention hierarchical transformer module (DAHT) is proposed to fully extract the dual-time feature maps, inspired by previous studies. In the decoding stage, a lightweight MLP decoder (LMLP) is proposed to fuse high-level and low-level features. The whole network structure is shown in Figure 1.Table 2 shows the full write-up and abbreviation of the content in the module of Figure 1.

As shown in Figure 1, a multi-scale approach is used to fuse the disparity features from high to low. The input dual-temporal images are downsampled to generate feature maps at different scales and resolutions. The difference maps of the two temporal phases at different scales are generated after processing by the transformer module. Finally, a convolutional network-like multilayer feature is generated with high-resolution coarse-grained features and low-resolution fine-grained features.

Specifically, given a pre-temporal or post-temporal image with resolution H $\times$ W $\times$ C, a feature mapping map with resolution F_i is output by the transformer module coding transformer encoder with resolution $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where $i = 1, 2, 3, 4$, $C_{i+1} > C_i$, this feature map is obtained by Difference Module, and then up-sampled by LMLP decoder to get the image with the same width and height as the input image, and finally the change map is obtained by classification.

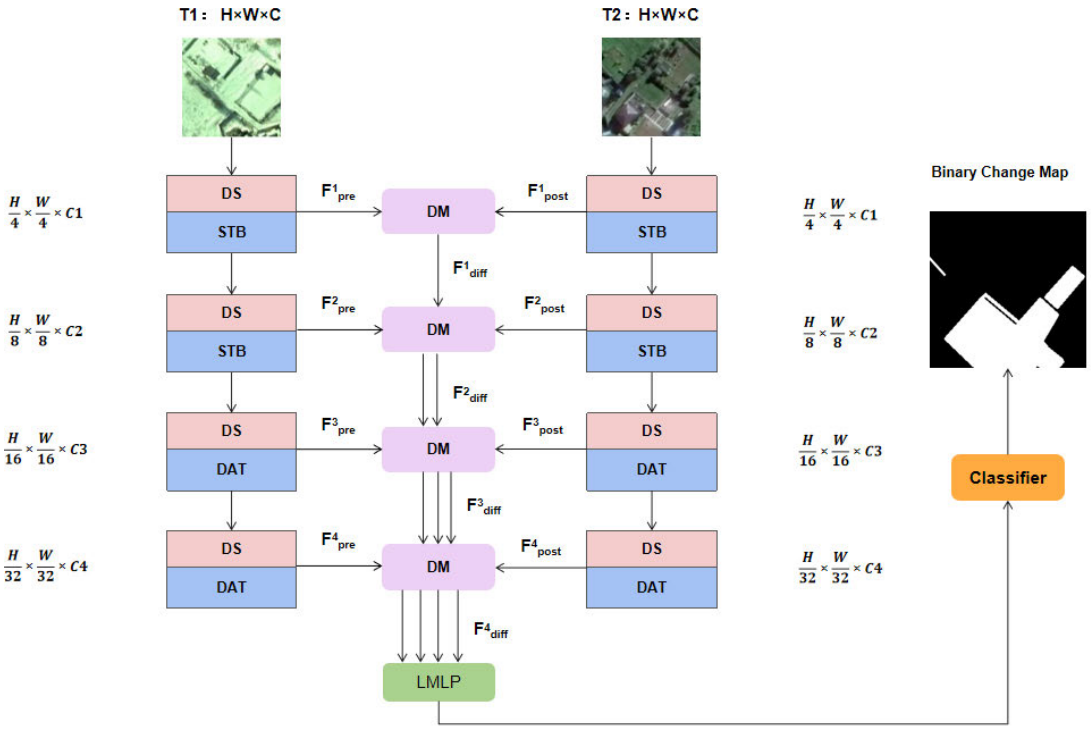


FIGURE 1. Architecture of the proposed DAHT-Net network.

TABLE 2. Table of Abbreviations.

Full Form	Abbreviation
DownSampling	DS
Swin Transformer Block	STB
Difference Module	DM
Deformable Attention Transformer	DAT
Lightweight Multi-layer Perceptron	LMLP

B. SWIN TRANSFORMER BLOCK(STB)

The transformer module consists of two consecutive swin transformer blocks as shown in Figure 2. Figure 2 is an example of the first layer.One Swin Transformer Block consists of a shifted window-based MSA with two layers of MLPs. LayerNorm (LN) layers are used before each MSA block and each MLP, and residual connections are used after each MSA and MLP.

First module W-MSA: Uses a regular window partitioning strategy starting from the top left pixel to uniformly divide the 8 × 8 feature map into 2 × 2 windows of size 4 × 4 (M = 4). The next module SW-MSA: uses a different window configuration from the previous layer by shifting the window by (M/2, M/2) pixels from the regularly divided window.

The 8 × 8 size feature map of the previous layer of the Swin Transformer Block has divided into 2 × 2 patches of size 4 × 4 each, and then the window positions of the next layer of Swin Transformer Block are shifted to obtain 3 × 3 non-overlapping patches. The shifting window is divided in such a way that the connection is introduced between the adjacent non-overlapping windows of the previous layer, which greatly increases the perceptual wilderness. You can see that the shifted window contains elements of the original neighboring windows.

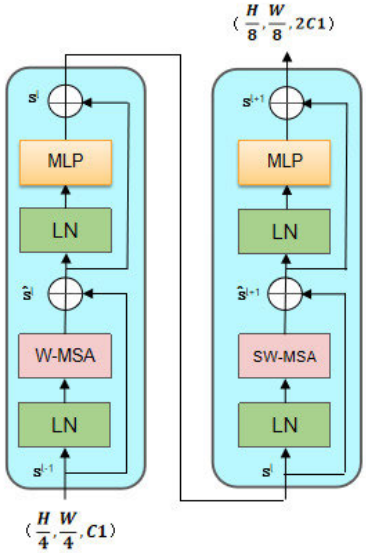


FIGURE 2. Architecture of the swin transformer block network.

But this also introduces a new problem, which is the increase in the number of windows from 4 to 9. This was achieved indirectly by shifting the feature map and setting the mask for Attention. The final result is equivalent while maintaining the original number of windows.

C. DEFORMABLE ATTENTION TRANSFORMER(DAT)

As shown in Figure 3, we first reduce the computational complexity by sequence normalization, then selectively focus on the small window where the context is located by local attention, and finally enter the transformer module for encoding.

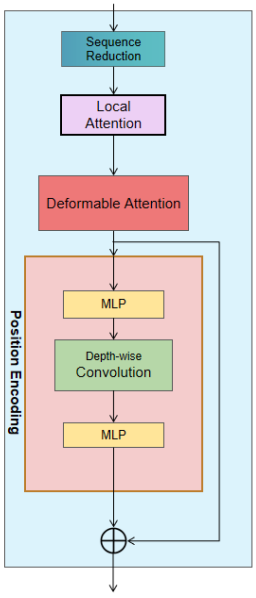


FIGURE 3. Architecture of the transformer block network.

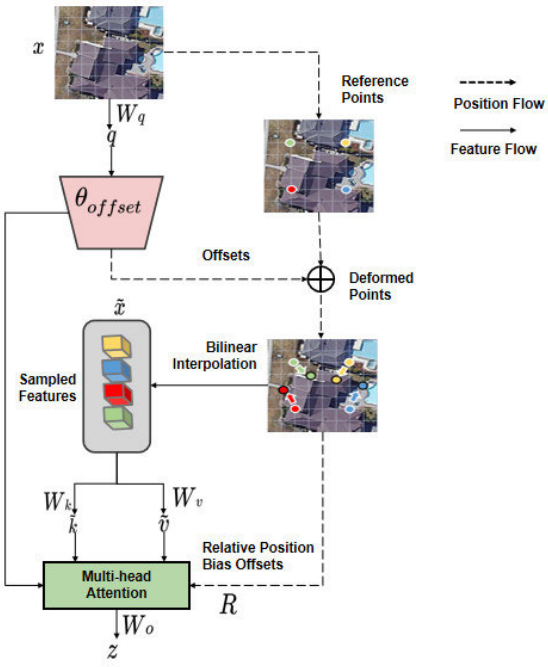


FIGURE 4. Deformable Attention-guided Hierarchical Transformer(DAHT).

Figure 4 shows the specific structure of the deformable attention for extracting global feature information in Figure 3.

Since the current manual sparse attention model [23], [26] leads to a large amount of information loss, the use of attention shifting leads to a slow growth of the receptive field, severely limiting the potential problem of modeling large objects. We propose deformable attention modules that are both flexible in terms of the set of candidate keys or values given a query, and produce different query results based on each individual input, inspired by Xia et al. [5].

As can be seen in the above figure, a set of reference points are placed uniformly on the feature map, and the offset network gets the offset by learning the query, and then projects the deformation points from the sampled features than the deformation values and keys. The relative position deviations of the deformation points are also calculated to enhance the multi-head attention and output the transformed features. We randomly choose 4 reference points for the representation, but there will be more points in the actual experiment.

To be precise, the deformable attention module effectively models the relationship between reference points by focusing on important regions in the feature map. These focused regions are determined by multiple sets of deformable sampling points, which are learned from the query by the offset network. A bilinear interpolation method is used to sample from the feature map, and then the sampled features are fed into the key and projection to obtain the deformed key and value. Finally, standard multi-headed attention is applied to engage the query on the sampled keys and aggregate features from the deformed values.

In addition, the location of the deformation points provides a more robust relative position bias to facilitate the learning of deformable attention. As shown in Figure 4, given the input feature map $x \in \mathbb{R}^{H\times W\times C}$, a uniform grid of points $p \in \mathbb{R}^{H_G\times W_G\times 2}$ are generated as a reference. Specifically, the grid size is downsampled by a factor r,$H_G = H/r$,$W_G = W/r$ from the input feature map size. The values of the reference points are linearly spaced two-dimensional coordinates $\{(0,0),\ldots,(H_G-1,W_G-1)\}$, which are then normalized to the range $[-1,+1]$, where $[-1,-1]$ denotes the upper left corner and $[1,1]$ denotes the upper right corner, according to the grid shape $H_G \times W_G$.

To obtain the offsets for each reference point, the feature mapping is linearly projected to obtain the query tokens $q = xW_q$, which are then fed into a lightweight subnetwork $\theta_{\text{offset}}(\cdot)$ to generate the offsets $\Delta p = \theta_{\text{offset}}(q)$. To stabilize the training process, we measure the amplitude of Δp by some predefined factor s to prevent the offset from becoming too large, i.e.,$\Delta p \leftarrow s\tanh(\Delta p)$. The features are then sampled at the locations of the deformation points as keys and values, followed by the projection matrix:

$$q = xW_q, \quad \tilde{k} = \tilde{x}W_k, \ \tilde{v} = \tilde{x}W_v, \tag{1}$$

$$\text{with } \Delta p = \theta_{\text{offset}}(q), \quad \tilde{x} = \phi(x; p+\Delta p) \tag{2}$$

k and $\tilde{v}$ denote the deformed key embedding and value embedding, respectively. Specifically, we set the sampling function $\phi(\cdot;\cdot)$ as a bilinear interpolation to make it differentiable:

$$\phi\left(z;\left(p_x,p_y\right)\right) = \sum_{(r_x,r_y)} g\left(p_x,r_x\right) g\left(p_y,r_y\right) z\left[r_y,r_x,:\right], \tag{3}$$

where $g(a,b) = \max(0, 1-|a-b|)$ and $\left(r_x, r_y\right)$ indexes all locations on. Since g is nonzero only at the 4 closest integration points, it simplifies Equation (8) to a weighted average of the 4 locations. Similar to existing methods, we perform multi-headed attention on q, k, and v with a relative position

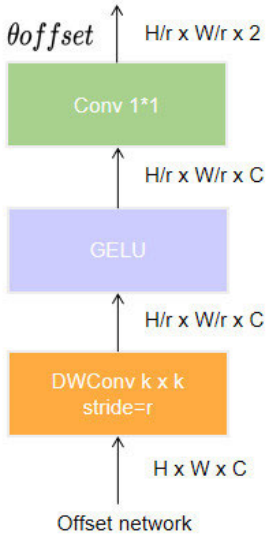


FIGURE 5. Offset network.

TABLE 3. Table of Abbreviations.

Full Form	Abbreviation
Convolution	Conv
Gaussian Error Linear Units	GELU
Depthwise Convolution	DWConv

offset R. The output of the attention head is formulated as follows:

$$z^{(m)} = \sigma\left(q^{(m)}\tilde{k}^{(m)\top}/\sqrt{d} + \phi(\hat{B}; R)\right)\tilde{v}^{(m)} \tag{4}$$

where $\phi(\hat{B}; R) \in \mathbb{R}^{HW\times H_G W_G}$ corresponds to the position embedding after the previous work [5], with some adaptation. Details will be explained later in this section. The features of each head are joined together and projected through Wo to obtain the final output z as Equation (3).

The structure of the bias network is shown in Figure 5. Table 3 shows the full and abbreviated contents of the module in Figure 5.

As described, the offset generation is performed using a sub-network, which is fed by the query features and the offset values of the output reference points, respectively. To ensure that the generative network learns reasonable offsets, it is important to consider that each reference point covers a local $s \times s$ region (where s represents the maximum value of the offset). Hence, the sub-network is implemented as two convolutional modules with nonlinear activation, as illustrated in Figure 5. The input features shown are initially obtained through a 5×5 deep convolution to capture local features.

Then, GELU activation and 1×1 convolution are used to obtain the two-dimensional offsets. It is also noteworthy that the bias of the 1×1 convolution is reduced to alleviate the forced offset at all locations.

To promote the diversity of deformation points, we follow a similar paradigm in MHSA and divide the feature channels into G groups. The features in each group generate the corresponding offsets separately using a shared sub-network. In practice, the number of heads M of the attention module is set to be a multiple of the size of the offset group G, ensuring that multiple attention heads are assigned to a set of deformed keys and values.

The relative position deviation encodes the relative position between each pair of queries and keys, which increases the common attention through spatial information. Consider a feature map of shape $H \times W$ with relative coordinate displacements in the ranges $[-H, H]$ and $[-W, W]$, respectively. In Swin Transformer [23], the relative position bias table $\hat{B} \in \mathbb{R}^{(2H-1)\times(2W-1)}$ is constructed and the relative position bias B is obtained by indexing in both directions.

Since the deformable attention in our approach involves continuous key positions, the relative displacements are computed within the normalized range of $[-1, +1]$. Subsequently, the interpolation $\phi(\hat{B}; R)$ is performed using the continuous relative biases in the parameterized bias table $\hat{B} \in \mathbb{R}^{(2H-1)\times(2W-1)}$, in order to encompass all possible offset values.

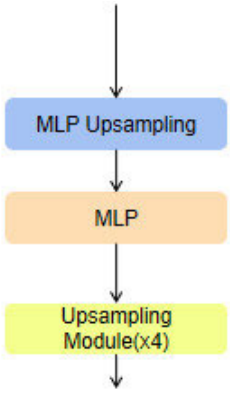


FIGURE 6. The structure of LMLP.

TABLE 4. Compare on LEVIR-CD dataset. All scores are expressed as percentages (%). The best scores are marked in red. Second place results are marked in blue.

Model	P	R	F1	IoU	OA
SNUNet	80.60	78.89	79.73	69.45	87.34
IFNet	87.86	73.94	80.30	62.96	87.83
STANet	89.86	81.68	85.69	87.66	97.49
BIT	89.75	86.18	87.92	89.78	93.41
STNet	90.06	89.03	89.54	82.09	99.36
Dsfer-Net	90.15	89.42	89.78	83.09	99.04
DAHT-Net	90.48	91.04	90.76	87.56	99.16

TABLE 5. Compare on CDD dataset. All scores are expressed as percentages (%). The best scores are marked in red. Second place results are marked in blue.

Model	P	R	F1	IoU	OA
SNUNet	89.18	87.17	88.16	78.83	98.04
IFNet	92.02	82.93	87.23	78.77	98.6
STANet	83.81	91.00	87.27	86.40	98.0
BIT	90.24	93.51	91.84	80.68	97.89
STNet	90.06	89.03	89.54	82.09	99.52
Dsfer-Net	90.98	91.51	91.24	91.04	98.54
DAHT-Net	94.25	93.46	93.85	93.43	99.3

D. LIGHTWEIGHT MULTILAYER PERCEPTRON(LMLP)

We use a simple decoder with an MLP layer to aggregate multi-level feature difference maps to predict change maps,

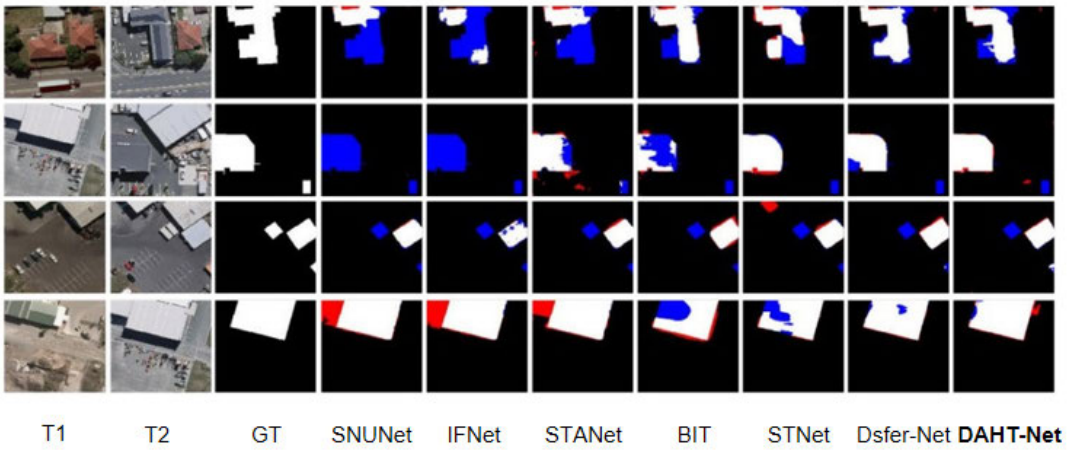


FIGURE 7. Description of the qualitative comparison of the dataset LEVIR-CD. The white color indicates the changes that were correctly detected. Black indicates that no changes have been correctly detected. Red indicates false alarms. Blue indicates unpredicted changes.

and its structure is shown in Figure 6. The proposed LMLP decoder consists of three main steps: MLP and upsampling, concatenation and fusion, and upsampling and classification.

First, the multiscale feature difference map is processed by the MLP layer in order to consolidate the channel dimensions. Next, each dimension is upsampled to a size of $\frac{H}{4} \times \frac{W}{4}$, as illustrated below.

$$\tilde{\mathbf{F}}^i_{\text{diff}} = \text{Linear}\ (C_i, C_{\text{ebd}})\left(\mathbf{F}^i_{\text{diff}}\right) \forall i, \tag{5}$$

$$\hat{\mathbf{F}}^i_{\text{diff}} = \text{Upsample}\ ((H/4, W/4), \text{“bilinear”})\left(\tilde{\mathbf{F}}^i_{\text{diff}}\right) \tag{6}$$

where C_{ebd} is the embedding dimension.The upsampled feature difference maps are connected and fused through an MLP layer as follows.

$$\mathbf{F} = \text{Linear}\,(4C_{\text{ebd}}, C_{\text{ebd}})\left(\text{cat}\left(\hat{\mathbf{F}}^1_{\text{diff}}, \hat{\mathbf{F}}^2_{\text{diff}}, \hat{\mathbf{F}}^3_{\text{diff}}, \hat{\mathbf{F}}^4_{\text{diff}}\right)\right) \tag{7}$$

A two-dimensional transposed convolutional layer is employed to upsample the fused feature map F to a size of $H \times W$, with $S = 4$ and $K = 3$. Finally, the up-sampled fused feature map is processed through another MLP layer to predict the change mask CM with a resolution of $H \times W \times N_{\text{cls}}$, where $N_{\text{cls}}\ (= 2)$ is the number of classes, i.e., change and no change. This process can be formulated as follows.

$$\hat{\mathbf{F}} = \text{ConvTranspose2D}\ (S = 4, K = 3)(\mathbf{F}) \tag{8}$$

$$\mathbf{CM} = \text{Linear}\ (C_{\text{ebd}}, N_{\text{cls}})\,(\hat{\mathbf{F}}) \tag{9}$$

E. LOSS FUNCTION

In the training stage, a cross-entropy loss function optimized by Chen and Shi [15] is used, which minimizes the cross-entropy loss to optimize the network parameters. Formally, the loss function is defined as Equation (10) [15]:

$$L = \frac{1}{H_0 \times W_0} \sum_{h=1,w=1}^{H,W} l\,(P_{hw}, Y_{hw}) \tag{10}$$

TABLE 6. Compare on WHU-CD dataset. All scores are expressed as percentages (%). The best scores are marked in red.Second place results are marked in blue.

Model	P	R	F1	IoU	OA
SNUNet	88.19	86.89	87.54	89.1	97.81
IFNet	90.54	84.59	87.45	87.5	98.42
STANet	87.59	91.52	89.52	89.8	99.0
BIT	90.62	89.35	89.98	92.1	99.6
STNet	87.84	87.08	87.46	77.72	98.85
Dsfer-Net	92.17	93.58	92.87	90.79	93.18
DAHT-Net	92.49	93.18	92.83	92.4	98.93

where $l\,(P_{hw}, y) = -\log\,(P_{hwy})$ is the cross-entropy loss and Y_{hw} is the label for the pixel at location (h, w) [15]

IV. EXPERIMENT

To validate the effectiveness of our proposed method, we compared the performance of separable transformers with state-of-the-art methods for image change detection. The image change detection and the comparison of the results with each method were performed on LEVIR-CD [27], CDD [28], and WHU-CD datasets [29], respectively. In addition, we conducted an ablation study to demonstrate the effectiveness of the Transformer basic module as well as variable attention. Our network is implemented on the PyTorch platform running on an NVIDIA Titan RTX 2080Ti with 11G RAM.

A. DATASETS AND EVALUATION METRICS

We evaluate the proposed DAHT-Net on three public datasets in five common metrics: precision (P), F1-score (F1), recall (R), intersection over union (IoU), and overall accuracy (OA).

LEVIR-CD [15] dataset contains 637 pairs of co-aligned very high resolution (VHR, 0.5m/pixel) Google Earth images, 1024 × 1024 pixels in size. The number of change pixels and constant pixels were 30,913,975 and 637,028,937, respectively. Due to GPU memory limitations, the raw images were cropped into smaller 512 × 512 pixel image blocks for model training and evaluation. In our case, the original image

is cropped into 16 image patches of size 256 × 256 pixels, generating 7120 pairs of image blocks for training, 1024 for validation, and 2048 pairs for testing.

CDD [16] dataset contains 11 pairs of diachronic images with seasonal variation, including 7 pairs of images of size 4725 × 2200 pixels and 4 pairs of images of size 1900 × 1000 pixels. In this paper, a subset of remotely sensed image data with seasonal variations is selected, and all images are segmented into 256 × 256 image patches by cropping and rotation to generate 16,000 pairs of patches. For these patches, 10,000 pairs are used for training, 3,000 pairs are used for validation, and the rest are used for testing.

The WHU-CD [29] dataset is a CD dataset for public buildings. It consists of a pair of HR (0.075 m) aerial images with dimensions of 32, 507 × 15, 354. The final choice was to crop the image into small blocks of size 224 × 224 and split it into three random sections: 7918/987/955 for training/validation/testing, respectively.

B. COMPARISON METHOD

To verify the effectiveness and superiority of our methods, we selected six methods represented in the CD task and compared the performance of these methods in CDD, LEVIR-CD, and WHU-CD, respectively, and a brief description of the selected methods is given below:

1) SNUNet [18] reduces the loss of location information in deep network training by tight information transfer between encoder and decoder, and also proposes an integrated channel attention mechanism (ECAM) for deep supervision.
2) IFNet [30] first used both Siamese network architectures as the original image feature extraction network. To improve the integrity of change map boundaries and internal density, an attention mechanism is used to fuse multi-level depth features with image difference map features.
3) STANet [15] used Siamese FCN for feature extraction and learned change maps based on the distance between dual temporal features. A new spatiotemporal attention neural network based on the bimodal network is proposed, which exploits the spatiotemporal dependence and designs a CD self-attentive mechanism to model spatiotemporal relationships. And a new HR remote sensing image dataset, LEVIR-CD, is proposed.
4) BIT [31] a transformer-based approach that represents the input image as some high-level semantic tokens. By adding a transformer encoder to the CNN backbone network, BIT-CD models the context in a compact token-based spacetime.
5) STNet [32] adopts a fusion of self-encoder and 3D CNN to enhance change detection while maintaining a lightweight and deployable model for various fields. However, extensive data is necessary for effective training, and the model structure is relatively intricate, requiring specialized technical expertise for design and implementation.
6) Dsfer-Net [33] adopts a combination of hierarchical feature extraction and attention mechanism, which enables efficient identification of change regions while retaining original feature information. The model contains a smaller number of parameters and runs at a faster speed, making it suitable for large-scale remote sensing image change detection. However, its accuracy may be slightly less than some models, such as Swin Transformer.

C. COMPARISON AND ANALYSIS

Firstly, we will first analyze our proposed DAHT-Net method with other methods by evaluating the metrics on each dataset, with red data in the table representing the best and blue data representing the second best. Secondly we also select two result plots of all methods from each of the three datasets and analyze them by direct observation. Where the red label represents false alarms, the blue labeled area represents unpredicted changes, and the white labeled area represents correct detection. Then we also performed ablation experiments to demonstrate the necessity and validity of each module of the model. Finally, we compare our number of parameters and detection speed with other methods in the form of a bar chart.

In this paper, we compare DAHT-Net with current state-of-the-art methods on three benchmark datasets. Table 4 and Figure 7 show the experimental results of all methods on the LEVIR-CD dataset along with the result plots. From Table 4, we can see that DAHT-Net achieves the best experimental results in terms of P-value, R-value and F1-value, which are 90.48%, 91.04% and 90.76%, respectively. Compared with other methods, our DAHT-Net predictions are more accurate and have higher capture ability. And the results in IoU and OA values are slightly lower than BIT and STNet network structures, respectively, but the difference is not significant.

Figure 7 shows the experimental results of the three images selected from the LEVIR-CD dataset, respectively, with the red region indicating the spurious region and the blue region indicating the missing region. From Figure 7, it can be seen that SNUNet, IFNet, STANet, and STNet methods have more false regions, and SNUNet, IFNet, and BIT methods have more missing regions. The buildings in the samples of these five methods are in close proximity to each other with unclear boundaries, which poses a challenge to the CD task.The Dsfer-Net method and our proposed DAHT-Net method have fewer false and missed regions, indicating that the building boundaries are more accurately recognized, and can accurately detect regions of obvious building changes.

The experimental results of all methods on the CDD dataset and the result plots are given in Table 5 and Figure 8. From Table 5, we can see that DAHT-Net has the best experimental results for P-value, R-value, F1-value and IoU-value, which are 94.25%, 93.46%, 93.85% and 93.43%, respectively. Compared with other methods, our DAHT-Net recognition is more accurate, and the predicted bounding box is more compatible with the real target location. As can be seen from Figure 8, the IFNet, STANet, BIT, and STNet

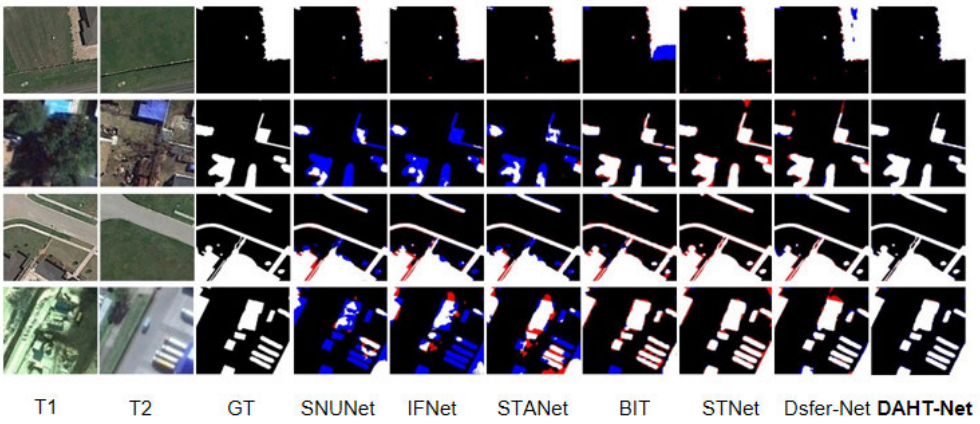


FIGURE 8. Description of the qualitative comparison of the dataset CDD. The white color indicates the changes that were correctly detected. Black indicates that no changes have been correctly detected. Red indicates false alarms. Blue indicates unpredicted changes.

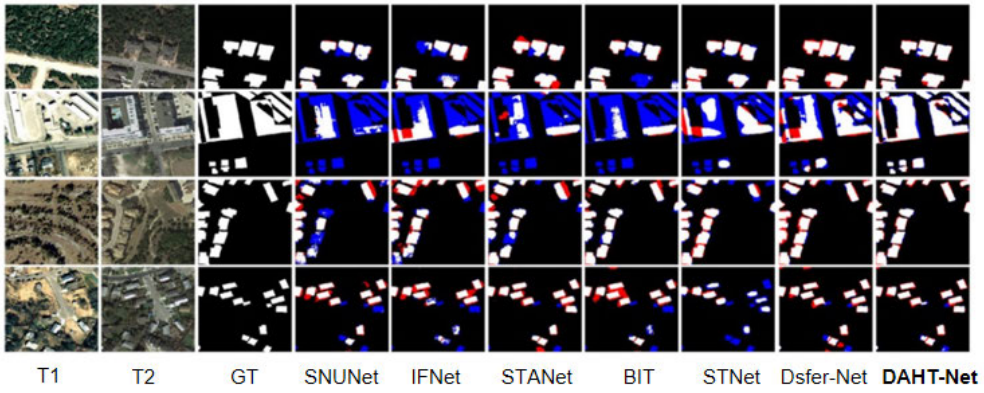


FIGURE 9. Description of the qualitative comparison of the dataset WHU-CD. The white color indicates the changes that were correctly detected. Black indicates that no changes have been correctly detected. Red indicates false alarms. Blue indicates unpredicted changes.

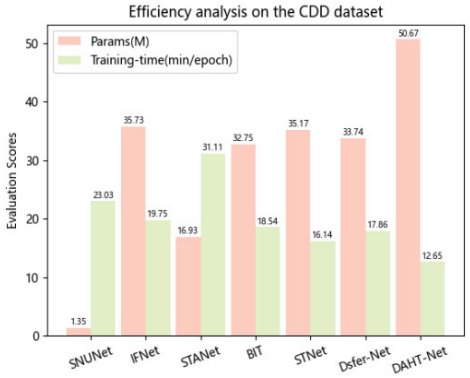


FIGURE 10. Illustration of an efficiency analysis of the comparison methods.

methods have more false regions, and the SNUNet, IFNet, and STANet methods have more missing regions. The edges of the samples of these methods are not clear, which leads to more missed regions and false regions in almost so methods. The DAHT-Net method, on the other hand, has the most accurate edge detection and more accurate detection of small change regions.

The experimental results of all methods on the WHU-CD dataset and the result plots are given in Table 6 and Figure 9.

TABLE 7. Ablation studies of different modules on CDD dataset. All scores are expressed as percentages (%). The best scores are marked in bold.

Model			CDD-CD		
Baseline	STB	DAT	F1	Kappa	OA
√	×	×	85.65	86.12	94.89
√	√	×	85.92	87.99	95.24
√	×	√	85.13	86.97	94.92
√	√	√	93.85	92.48	99.3

From Table 6, we can see that DAHT-Net has the best experimental results in P-value and IoU-value, which are 92.49% and 92.4%, respectively.The R-value and F1-value are slightly lower than the Dsfer-Net method, and the OA-value is slightly lower than that of the BIT method.

The performance of DAHT method is not very good on WHU-CD dataset, which is considered to be due to the unbalanced distribution of the samples in WHU-CD dataset. balanced distribution of samples in the WHU-CD dataset. From Figure 9, it can be seen that the DAHT-Net method has more false regions, except for the DAHT-Net method, which has more missed regions. Overall the DAHT-Net method is able to extract more effective semantic information and has a better detection effect.

In addition, we use the number of parameters in millions (called Params (M)) and the number of floating point

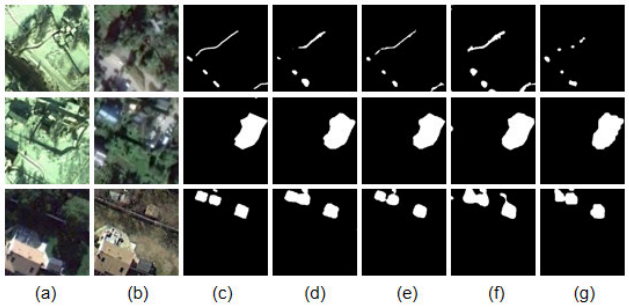


FIGURE 11. Visualization comparison plots of each network in the ablation experiment on CDD datasets. (a) Image T1.(b) Image T2. (c) Field facts. (d) Baseline. (e) Baseline+STB. (f) Baseline+DAT. (g) Baseline+STB+DAT.

operations per second in gigabytes (called Flops (G)) to measure the space complexity and computational cost of our DAHT-Net with some comparison methods. The horizontal axis is the name of the method to be compared, the vertical axis is the evaluation score value, the pink bar represents the number of parameters, and the green bar represents the number of floating point operations per second, as shown in Figure 10.

The method requires more parameters i.e. 50.67M and has the lowest computational cost of 12.65 min/epoch, which is superior to other change detection methods in terms of computational cost.Although the proposed network DAHT-Net achieves encouraging performance, it has some potential limitations. The computational complexity of DAHT-Net is relatively high and the number of parameters is large. This is not friendly to devices and applications with limited resources. However, from another perspective, the training efficiency of the proposed DAHT-Net is also relatively impressive. Compared with STANet and SNUNet, the training time of the proposed method is reduced by 59.34% and 45.10%, respectively, which makes the proposed method more valuable in practical applications under the same equipment conditions. Though the number of training parameters and training time are comprehensive, the proposed method has space for improvement and enhancement in the future. For example, model compression can be performed in the proposed network, employing pruning and knowledge distillation [34], [35] to reduce the size of the model.

In addition to this, we conducted an ablation study of DAHT-Net on the CDD dataset. Specifically, we chose a variant of DAHT-Net (i.e., without STB and DAT modules) as the baseline model. As shown in Table 7, the introduction of the STB module has improved its performance. However, with the introduction of the DAT module, there is a decrease in the F1 value, but an improvement in other metrics. After the introduction of STB module and DAT module, there is a significant improvement in their performance which proves their correctness. Among them, DAHT-Net has the highest F1 of 94.25%, which indicates that the combination of STB module and DAT module can significantly improve the change detection performance.

Figure 11 shows the resulting plots for each of the three images selected from the CDD dataset. As can be seen from the figure, adding STB and DAT modules from the baseline model is much more accurate than adding only STB modules or DAT modules. The baseline model, on the other hand, misses the detection. This shows that there is a need for each of our modules.

V. CONCLUSION

In this paper, a novel hierarchical transformer network architecture based on deformable attention is proposed for remote sensing image change detection. The first two layers used in this paper use Swin transformer blocks and the last two layers use Deformable Attention Transformer (DAT) modules to extract multi-scale features. Among them, the DAT module is able to extract better feature representations and make the model more sensitive to changes by dynamically focusing on the region of interest through adaptive attention. Finally, the extracted multiscale features are quickly recovered to a variation map with the same width and height as the original map of dimension 2 by the LMLP module. In this paper, we conducted experiments on three datasets, CDD, LEVIR-CD and WHU-CD, to demonstrate the effectiveness of the DAHT-Net method. In addition, DAHT-Net has the advantages of reducing the amount of data, enhancing the feature representation and improving the model robustness.

However, the DAHT-Net method does not perform well on the WHU-CD dataset compared to Dsfer-Net. Considering the relatively high image resolution of the WHU-CD dataset, there is more noise and background interference, which leads to a lot of leakage in the change detection effect. Therefore, in the future, we plan to extend our work to the task of image change detection based on generative adversarial networks.

REFERENCES

[1] C. Zhang, Y. Zhang, and H. Lin, "Multi-scale feature interaction network for remote sensing change detection," *Remote Sens.*, vol. 15, no. 11, p. 2880, Jun. 2023.

[2] C. Ren, X. Wang, J. Gao, X. Zhou, and H. Chen, "Unsupervised change detection in satellite images with generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10047–10061, Dec. 2021.

[3] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, Sep. 2021.

[4] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.

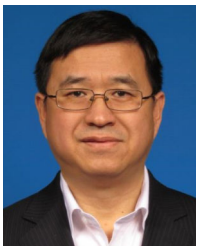
[5] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4784–4793.

[6] 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.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf., Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[8] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.

[9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.
[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[11] 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.
[12] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.
[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.–MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[14] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.
[15] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.
[16] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "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.
[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. (CVPR)*, Jul. 2017, pp. 2261–2269.
[18] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[19] 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.
[20] L. O. Chua and T. Roska, "The CNN paradigm," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 40, no. 3, pp. 147–156, Mar. 1993.
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[22] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.
[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[24] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[25] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9300–9308.
[26] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.
[27] Y. Zhang, S. Zhang, Y. Li, and Y. Zhang, "Coarse-to-fine satellite images change detection framework via boundary-aware attentive network," *Sensors*, vol. 20, no. 23, p. 6735, Nov. 2020.
[28] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 565–571, May 2018.
[29] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[30] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.
[31] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.
[32] X. Ma, J. Yang, T. Hong, M. Ma, Z. Zhao, T. Feng, and W. Zhang, "STNet: Spatial and temporal feature fusion network for change detection in remote sensing images," 2023, *arXiv:2304.11422*.
[33] S. Chang, M. Kopp, and P. Ghamisi, "Dsfer-Net: A deep supervision and feature retrieval network for bitemporal change detection using modern Hopfield networks," 2023, *arXiv:2304.01101*.
[34] M. A. Carreira-Perpinan and Y. Idelbayev, "'Learning-compression' algorithms for neural net pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8532–8541.
[35] M. P. Vadera and B. M. Marlin, "Challenges and opportunities in approximate Bayesian deep learning for intelligent IoT systems," in *Proc. IEEE 3rd Int. Conf. Cognit. Mach. Intell. (CogMI)*, Dec. 2021, pp. 252–261.



GANG SHI received the degree in control engineering from the Dalian University of Technology, Dalian, China, and the Ph.D. degree in computer science and technology from Tsinghua University. From 2015 to 2017, he visited the University of Minnesota System, USA, as an Exchange Ph.D. Student. Since 2017, he has been the Deputy Head of the Department of Computer Science and Technology, School of Information Science and Engineering, Xinjiang University. His research interests include artificial intelligence, the Internet of Things, and big data analytics. More detailed information about him can be found at http://it.xju.edu.cn/info/1144/2113.htm.

YUNFEI MEI received the bachelor's degree in computer science and technology from Xinjiang University, in 2020, where she is currently pursuing the master's degree in computer technology. Her research interests include deep learning, remote sensing, image change detection, and computer vision.

XIAOLI WANG received the M.S. degree from Xinjiang University, China. Since 2000, she has been an Associate Professor with the College of Information Science and Engineering, Xinjiang University. She has led several natural science foundation projects in this field, e.g., research on super-resolution reconstruction of remote sensing images based on deep learning of time-space-spectral features. Her research interests include machine learning and computer vision. More detailed information about her can be found at http://it.xju.edu.cn/info/1144/1703.htm.

QINGWEN YANG is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences. She is a Lecturer with the School of Information Science and Engineering, Xinjiang University. Her research interest includes computer vision.