Change Detection in Remote-Sensing Images Using Pyramid Pooling Dynamic Sparse Attention Network With Difference Enhancement

Zhong Li, Bin Ouyang, Shaohua Qiu, Xinghua Xu, Xiaopeng Cui, and Xia Hua

***Abstract*—Benefits from the powerful local modeling capability of deep convolutional neural networks (CNNs), remote-sensing image change detection (CD) has made significant progress. In recent years, the rise of transformers has further driven improvements in global feature extraction for bitemporal remote-sensing images. Some prior efforts have tried to integrate CNN and transformer, but they suffer from the limitation of inefficiently aggregating local features and contextual information. Besides, they struggle to refine change boundaries and exhibit inferior performance in detecting multiscale and subtle changes. To tackle these interrelated problems, we propose a difference-enhanced pyramid pooling dynamic sparse attention network (DPDANet) for CD, which integrates the potential of CNN and pyramid pooling dynamic sparse attention (PDSA) mechanism. Specifically, a pretrained EfficientNetV2-S network is first used to extract multilevel local fine-grained features. Then, a global semantic enhancement network based on well-designed PDSA mechanism is proposed to extract rich global contextual information. The proposed difference enhancement module combines long short-term memory and deformable convolution to emphasize relevant and irrelevant changes, capturing precise boundary details of the changing region. A decoder is then employed for step-by-step upsampling of encoded features, with skip connections between local multiscale features and globally enhanced features. Expensive experiments on four public CD datasets demonstrate that DPDANet outperforms state-of-the-art methods by reducing missed detections and false detections and achieving more accurate boundaries of the changing area.**

***Index Terms*—Attention mechanism, change detection (CD), deep learning, remote sensing.**

I. Introduction

REMOTE-SENSING image change detection (CD) aims to achieve Earth observation by analyzing bitemporal images from the same scene and extracting regions of interest changes. It is widely used in fields, such as environmental surveillance [1], agricultural monitoring [2], [3], urban planning [4], [5], and disaster assessment [6].

The current remote-sensing image CD tasks face several common challenges. First, changes in lighting and seasons can cause nonuniform spectral features on the object's surface, leading to uncertainty in CD. In addition, scenes may have multiple scale changes, and small target changes may be easily drowned out by noise, making it urgent to improve the ability to extract multiscale features while ensuring robustness and generalization ability. Finally, the number of pixels in changing areas is typically less than that in invariant areas, resulting in an imbalance problem that needs to be addressed to improve detection accuracy.

Traditional CD methods primarily utilize spectral information from remote-sensing images to identify changes, which can be divided into two main types: 1) pixel-based [7], [8], [9] and 2) object-based [10]. Specifically, pixel-based methods generate change maps through a direct comparison of pixel values from multitemporal images and subsequent division by a predefined threshold. However, this type of method only considers the spectral changes of a single pixel and ignores the contextual information. As a result, it requires extremely high registration accuracy and empirical threshold division and is easily affected by environmental noise. Object-based methods, on the other hand, focus on both spectral and spatial information. However, due to the uncertainty of the changing object, its classification error is significant, which affects the detection accuracy. It is clear that conventional methods of CD based on pixels or objects have restricted feature extraction capabilities and are not fully efficient in leveraging the information contained in high-resolution images. Moreover, both of these methods require significant manual intervention and are more vulnerable to external noise.

With the continuous advancement of deep learning technology, convolutional neural networks (CNNs) have been successfully utilized in CD tasks. These works mainly employ CNN to extract deep features from remote-sensing images, which encompass abundant spectral and spatial details. These features are then used to establish semantic feature descriptions and achieve the detection of interested changes. Zhang et al. [11] proposed a spatial logical aggregation network based on morphological transformations, which utilizes spatial

Manuscript received 9 November 2023; revised 22 January 2024 and 22 February 2024; accepted 4 March 2024. Date of publication 6 March 2024; date of current version 29 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62102436 and Grant 62301587, in part by the Natural Science Foundation of Hubei Provincial under under Grant 2020CFB339 and Grant 2022CFB989, in part by the Foundation for National Key Laboratory of Science and Technology under Grant 6142217210503, and in part by the Project Foundation of University (NUE) under Grant 202250E050. *(Corresponding author: Shaohua Qiu.)*

Zhong Li, Bin Ouyang, Xinghua Xu, Xiaopeng Cui, and Xia Hua are with the National Key Laboratory of Electromagnetic Energy, Naval University of Engineering, Wuhan 430030, China.

Shaohua Qiu is with the National Key Laboratory of Electromagnetic Energy, Naval University of Engineering, Wuhan 430030, China, and also with East Lake Laboratory, Wuhan 430202, China (e-mail: qiush125@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3374050

morphological differences to enhance fine-grained boundaries. Li et al. [12] introduced a graph feature extraction module to extract topological structural information, which is then combined with rich spectral and spatial information. Wang et al. [13] proposed a multiscale interactive fusion network that utilizes a multiscale interactive information extraction block to extract rich scale information, and a global dependence fusion module to capture long-term dependencies. Wang et al. [14] proposed a multistage self-guided separation network. It addresses the interclass sample similarity issue caused by target-background imbalance through target-background separation strategy and contrastive regularization. Nevertheless, owing to the limited receptive field of convolution, CNN lacks the ability to model long-range spatiotemporal dependence and global context, making it susceptible to various noise changes. To address this limitation, CNN continues to evolve toward deeper and more complex structures. Some works have introduced spatial, channel, and multifrequency attention [15], [16] mechanisms to compensate for the shortcomings of global relationship modeling. Although these methods have achieved impressive results, they have also resulted in a significant increase in parameters. With the powerful ability of transformers to model global contexts and capture long-range dependencies, outstanding performance has been demonstrated by vision transformers in CD tasks. However, one main issue with transformer-based methods is that the boundaries of the generated change maps are relatively blurry, which may be caused by the loss of spatial information during token embedding and feature reconstruction [17]. Moreover, applying a transformer to remote-sensing image feature mapping can be computationally expensive due to the attention operation of the self-attention (SA) mechanism on the flattened feature vectors.

To alleviate the problems of high computational complexity and large memory usage, Li et al. [18] developed a lightweight and efficient CD model by optimizing the network structure and employing the lightweight backbone network MobileNetV2 [19] instead of the commonly used cumbersome backbone, i.e., VGG [20] and ResNet [21]. Zhang et al. [22] introduced CycleMLP [23] into the CD field, abandoning traditional convolution and SA operations. The introduction of CycleMLP brings about a computational complexity that is linearly correlated with the input size, offering a more efficient strategy. For transformer-based CD methods, numerous scholars have advocated for the incorporation of different levels of sparsity into the SA mechanism. This approach aims to direct each query to concentrate on a restricted set of key–value pairs, thereby mitigating computational complexity. Zhang et al. [24] employed the Swin transformer [25] block as the basic unit to construct a CD model. They utilized a shift window strategy to calculate the SA of local windows, effectively reducing the computational complexity associated with SA. Currently, several additional sparse patterns have been proposed, including static sparse patterns such as dilated windows [26] and axial stripes [27], as well as dynamic sparse patterns such as bilevel routing [28]. These sparse attention mechanisms enable efficient and effective modeling of long-range spatiotemporal dependencies while maintaining the spatial information inherent in the input image.

To enhance model accuracy and facilitate practical deployment, this article proposes an end-to-end encoding–decoding CD model [i.e., difference-enhanced pyramid pooling dynamic sparse attention network (DPDANet)] from the perspective of the effectiveness of feature extraction and integration, as well as reducing computational overhead, by combining EfficientNetV2-S and pyramid pooling dynamic sparse attention (PDSA). Fang et al. [29] demonstrated the significance of low-level features in capturing detailed spatial information through experiments. However, traditional backbone networks such as ResNet and UNet use plain convolutions to produce regular reception fields. Limited by the characteristics of the network, it is easy to lose low-level detail information during the feature extraction process. We propose using pretrained EfficientNetV2-S as the backbone network to alleviate this problem, which can enhance feature extraction efficiency and mitigate the limitations associated with inadequate training data [30]. In addition, to refine the fine-grained segmentation results of different semantics and enhance change information, we propose a difference enhancement module (DEM) that imposes constraints on the change boundaries. Inspired by the dynamic perception sparse attention mechanism proposed by Zhu et al. [28], we propose a global semantic enhancement module (GSEM) based on the designed PDSA that achieves dynamic sparsity from coarse to fine. It is worth noting that the objects of our dynamic sparse encoding are key–value pairs after pyramid pooling at different scales. Through pyramid pooling operations, the encoded feature sequence of the backbone is compressed, thereby reducing the memory occupation of subsequent dynamic sparse encoding while reducing the computational load of the model and capturing highly abstract multiscale information. In summary, the main contributions of this article can be summarized as follows.

1) We propose an end-to-end network architecture that ingeniously integrates the powerful local feature extraction capability of EfficientNetV2-S with our proposed PDSA mechanism, enabling efficient modeling of global contextual information. This architecture enables comprehensive learning of multiscale features and global semantic relationships.
2) We propose a GSEM based on PDSA, which can significantly reduce the computational load of plain multihead SA (MHSA) while extracting more representative multiscale contextual information and mitigating the impact of noise changes.
3) To tackle the issue of blurred boundaries between distinct semantics, we propose a DEM. It maximizes the difference between semantic changes of interest and noise changes to obtain precise boundaries of change regions, thereby improving the detection ability for small target changes.
4) The experimental results demonstrate that our DPDANet achieves state-of-the-art (SOTA) performance on four public datasets for remote-sensing image CD. It can effectively detect the changing regions of interest and obtain

more accurate boundaries, which are significantly superior to other benchmark models.

The rest of this article is organized as follows. Section II reviews the relevant work in the field of CD. Section III introduces the proposed PDSANet method in detail. Section IV reports on the experiment and analysis of the results. Section V presents the discussion. Finally, Section VI concludes this article.

II. Related Work

A. CNN-Based CD Methods

With the powerful feature representation capabilities of CNN, remote-sensing image CD methods based on CNN have become increasingly popular in recent years. These methods typically enhance the semantic representation of the network through changes to the network structure, optimization of the loss function, and the addition of attention mechanisms. Du et al. [31] proposed a bilateral semantic fusion twin network that integrates shallow and deep semantic features to obtain complete and refined boundaries of changing regions. Zhang et al. [32] replaced traditional convolution operations with dilated convolution [33], which increased the receptive field and improved the effect of CD. UNet++ [34] uses dense skip connections to extract multiscale features and mitigate pseudochanges. Li et al. [18] proposed a lightweight network based on progressive feature fusion and supervised attention, achieving efficient aggregation of multilevel features through designed neighbor aggregation, progressive change identification, and supervised attention modules. To further improve the performance of CNN models, some works have adopted more complex backbone networks, such as ResNet18 [35] and ResNet50 [36]. In addition, attention mechanisms have been introduced into CD tasks to enhance the features of changes of interest and weaken invariant features, including spatial attention, channel attention, and position attention. Fang et al. [37] proposed a channel attention module to fuse features at different levels. Peng et al. [38] incorporated dense attention connections between features from different layers to extract more comprehensive and impactful features. Chen et al. [39] proposed a pyramid spatiotemporal attention model to improve network detection performance by extracting features at different scales.

The aforementioned methods have improved the accuracy of CD to some extent. Nevertheless, the receptive field of CNN is constrained by its convolution operation, which only considers its connected image pixels, thus limiting its capability to extract global semantic features. Consequently, CNN-based methods are vulnerable to noise changes such as lighting and shadows, leading to reduced effectiveness of the semantic features. Despite the introduction of attention mechanisms to enhance the features of channel or spatial scales, the ability to learn global contextual relationships remains constrained, with a high computational complexity.

B. Transformer-Based CD Methods

The transformer model was initially applied in the field of natural language processing (NLP) [40] and later extended to image classification [41], image recognition [42], semantic segmentation [43], and other fields, achieving performance comparable to or even beyond the CNN model. Unlike CNN, transformer employs stacked MHSA modules to simulate the global relationship between labeled image blocks. ViT [44] was the first pure transformer network used for computer vision. Chen et al. [45] first introduced a transformer into the field of image CD. They proposed a bitemporal image transformer (BIT) that achieves CD through contextual modeling within a spatiotemporal range. While the sequential processing of BIT is beneficial for ensuring efficiency, it does not consider direct information exchange between CNN and transformer. In addition, obtaining changes in the original resolution directly through upsampling can easily lead to the omission of fine-grained details. Feng et al. [46] proposed a network based on intrascale crossover and interscale feature fusion, integrating CNN and transformer in parallel to enhance the synergy between local and global features. However, the parallelized structure requires a large amount of resource consumption. Zhang et al. [24] proposed a Siamese U-shaped transformer network for CD. They used a pretrained swin transformer on the ImageNet dataset to initialize the parameters of the proposed model, which further improved the performance.

ViT demands high computational resources, and its complexity increases quadratically with a size of input features or images. To lower the computational complexity, Li et al. [18] implemented a moving window strategy to compute SA within local windows, progressively expanding the network's receptive field through layer stacking. In addition, sparse patterns have been integrated into the SA mechanism by certain researchers, including dilated windows [26] and cross-shaped windows [27]. PVT [47] and MViT [48] used a single pooling operation in the MHSA module to downsample the feature maps, using the pooled feature to simulate the token-to-region relationship. Drawing inspiration from the work in [28], we implement dynamic sparsity from coarse to fine. It is worth noting that the objects of our dynamic sparse encoding are key–value pairs after pyramid pooling at different scales. The pyramid pooling operation reduces the memory storage of dynamic sparse encoding and introduces highly abstract multiscale information.

III. Proposed Method

A. Network Architecture

As illustrated in Fig. 1, the proposedDPDANet consists of three parts, including 1) feature extraction, 2) multiscale feature aggregation, and 3) prediction head. Given bitemporal images with a spatial resolution of 256×256 and channel number of 3. DPDANet first uses weight-shared EfficientNetV2-S as a feature extractor to extract multiscale local detail features. Then, the feature sequences of the bitemporal images are concatenated together, and the global semantic relationship is modeled using a GSEM to obtain contextual rich global features. Next, the multiscale low-level features are refined through the designed DEM to improve the distinguishability of changing features. Subsequently, advanced semantic features and enhanced low-level features of different scales are gradually fused. Finally, a

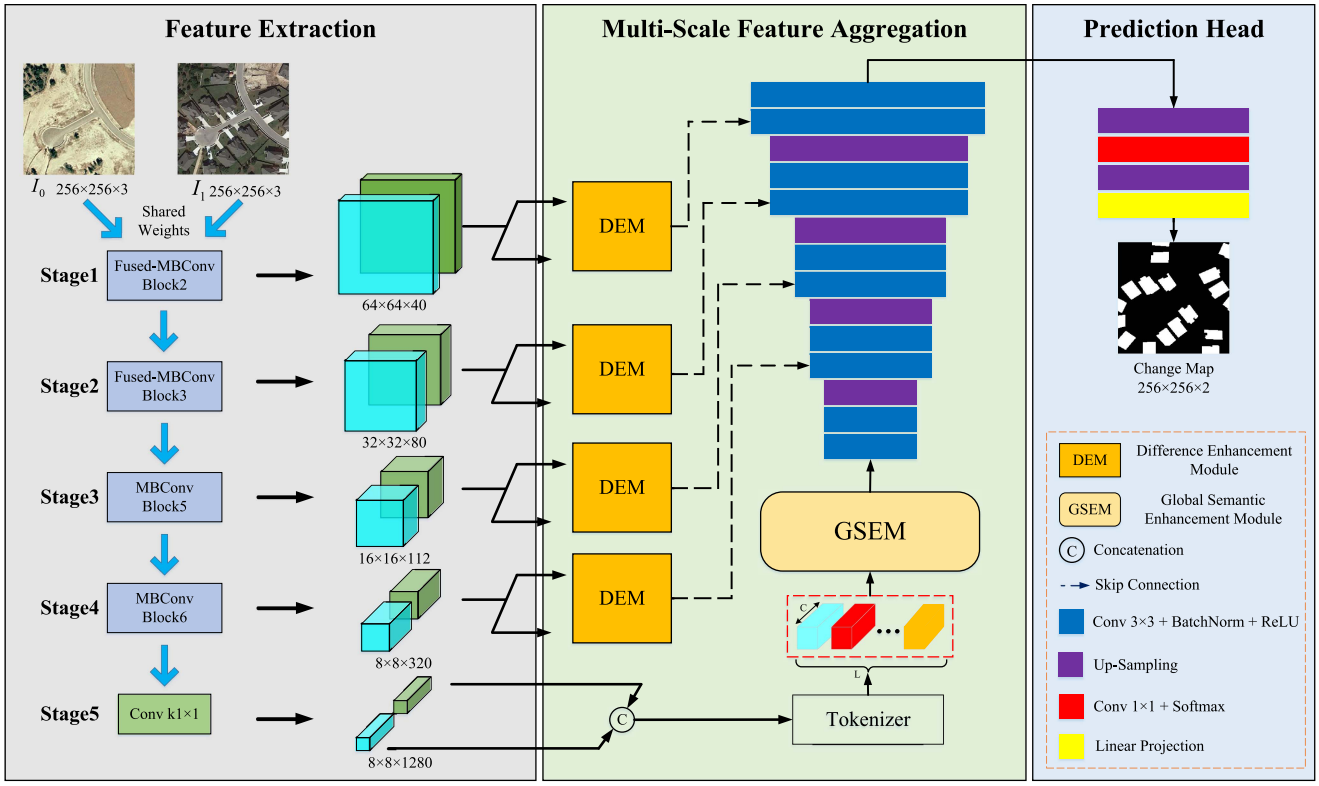

Fig. 1. Overall architecture of the proposed DPDANet model.

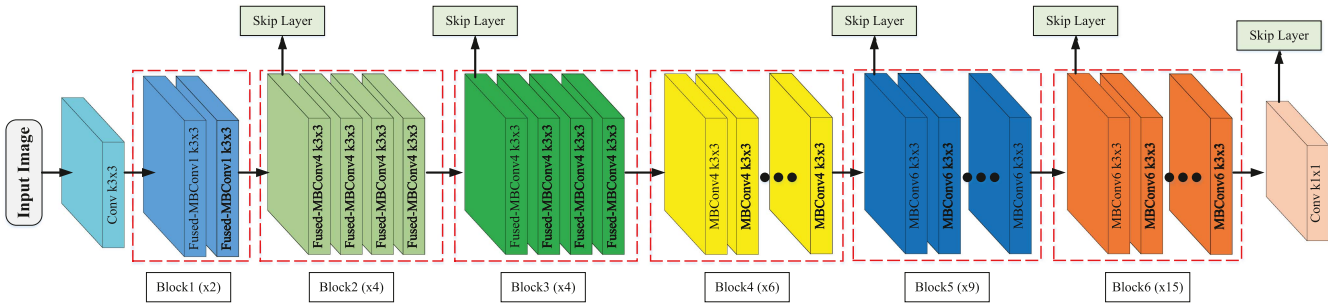

Fig. 2. Architecture of EfficientNetV2-S.

lightweight CNN is utilized to remap the refined features back to the original pixel space dimensions, generating a binary change map.

B. Feature Extraction Based on EfficientNetV2-S

In image CD tasks, cumbersome CNN backbone networks such as ResNet [35], [36], UNet [49], VGG19 [20], and Xception [50] are commonly used as feature extractors. Google researchers have introduced a more efficient CNN model, named EfficientNetV2. This model is uniformly adjusted based on the depth, width, and image resolution of the network, without the need for complex manual parameter tuning. It achieves better performance than other benchmarks on ImageNet and has fewer parameters. The building blocks of EfficientNetV2 consist of shallow fused mobile inversed bottleneck convolution (Fused-MBConv) modules and deep mobile inversed bottleneck convolution (MBConv) [19] modules, which have fewer parameters than traditional convolutional layers. In addition, the squeeze and excitation (SE) module of MBConv can enhance the detection ability of small target changes. We employ EfficientNetV2-S as the backbone network to extract multiscale local features and accelerate the training process through the weight of pretraining.

The architecture of EfficientNetV2-S is presented in Fig. 2. It first uses a 3×3 convolutional layer for preliminary feature extraction. Then, efficient feature extraction is achieved by repeatedly stacking Fused-MBConv modules and MBConv modules. Finally, it uses a 1×1 convolutional layer to output the final deep feature. Considering that different levels of the network can capture details and structural features of images at different scales, thus enhancing the model's adaptability to

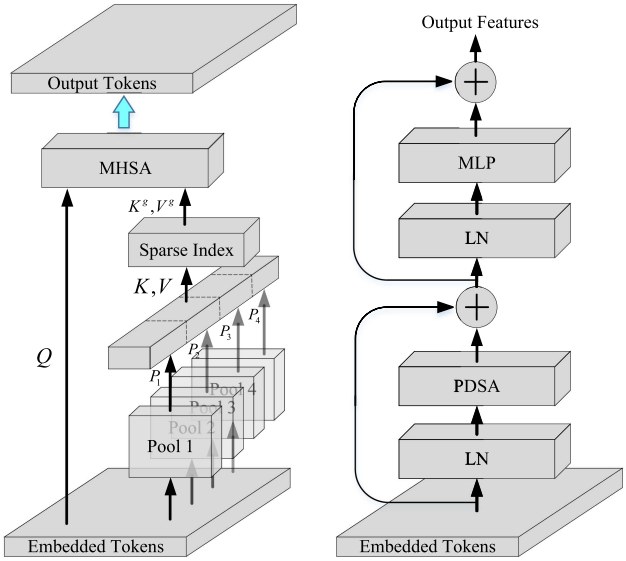

Fig. 3. *Left*: Details of the PDSA module. *Right*: The overall architecture of the GSEM.

objects of varying sizes and shapes in remote-sensing images, we aggregate multiscale features to improve the model's robustness in complex scenes [51], [52], [53]. Specifically, we use the first layer of Block2, Block3, Block5, and Block6 as the skip layer to produce multiscale features. Subsequently, the output features from the final convolutional layer undergo encoding via the GSEM.

C. Global Semantic Enhancement Module

The multiscale local features extracted by CNN lack global semantic information, which can limit their effectiveness in certain applications. For instance, in remote-sensing images CD tasks, misclassification can occur due to information redundancy. To tackle this issue, we propose a novel approach that incorporates pyramid pooling and dynamic sparsity into the remote-sensing image CD task. By utilizing a carefully designed PDSA module to construct GSEM, we aim to learn more efficient contextual information while reducing the computational complexity of global SA and improving the computational efficiency of the model.

1) PDSA Module: The structure of PDSA is shown in Fig. 3(a). Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, we first divide it into $S \times S$ nonoverlapped patches to achieve reshaping. Then, a pyramid feature map is generated by applying four average pooling layers with varying pooling ratios (empirically set as [12, 16, 20, 24] in this article) on the reshaped X'. This procedure can be defined as follows:

$$P_i = \text{AvgPool}_i(X'),\ i = 1, 2, 3, 4 \tag{1}$$

where P_i denotes the generated pyramid feature map. i represents the quantity of pooling layers. Subsequently, the pyramid feature map is fed into depthwise convolution [54] for relative position encoding. This procedure can be defined as follows:

$$P_i^{\text{enc}} = \text{DWConv}(P_i) + P_i,\ i = 1, 2, 3, 4 \tag{2}$$

where $\text{DWConv}(\cdot)$ denotes depthwise convolution, with a kernel size of 3×3. P_i^{enc} denotes P_i after relative position encoding. Next, these pyramid feature maps were flattened and concatenated together. This procedure can be defined as follows:

$$P = \text{LN}(\text{Concat}(f(P_i^{\text{enc}}))),\ i = 1, 2, 3, 4 \tag{3}$$

where f denotes the flattening operation. Concat denotes series connection. LN denotes layer normalization [55].

Dynamic sparsity is sought through the construction of a directed graph. Specifically, region-level queries and keys, $Q^r, K^r \in \mathbb{R}^{S^2\times C}$, are derived by applying per-region average on Q and K, respectively. We then calculate the matrix multiplication between Q^r and K^r to get the adjacency matrix $M^r \in \mathbb{R}^{S^2\times S^2}$. This procedure can be defined as follows:

$$M^r = Q^r (K^r)^T. \tag{4}$$

The adjacency matrix shows the degree of correlation between different regions. On this basis, we use the rowwise topk operator to achieve pruning, retaining only the first k connections of each region (k defaults to 4) and obtain the index matrix $I^r \in \mathbb{R}^{S^2\times k}$

$$I^r = \text{topk}(M^r). \tag{5}$$

Therefore, for each query token in the region, it only focuses on the corresponding $I^r_{(i,1)}, I^r_{(i,2)}, \ldots, I^r_{(i,k)}$. We extract the corresponding key and value tensor of interest through the gather operator. This procedure can be defined as follows:

$$K^g = \text{gather}(PW^k, I^r) \tag{6}$$

$$V^g = \text{gather}(PW^v, I^r) \tag{7}$$

where W^k and W^v denote the weight matrix of linear transformations that generate K and V, respectively. $K^g, V^g \in \mathbb{R}^{S^2\times \frac{kHW}{S^2}\times C}$ denote the gathered key and value tensor of interest after dynamic sparse encoding, respectively. Thus, in the MHSA module, the tensor calculation for query (Q), key (K), and value (V) is transformed as follows:

$$(Q, K, V) = (X'W^q, K^g, V^g) \tag{8}$$

where W^q denotes the weight matrix of the linear transformations that generates Q. Finally, we further calculate the correlation between Q and K by applying dot product operation and Softmax activation, which is used as the weight of V to calculate attention feature A. This procedure can be defined as follows:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

where d_k denotes the dimension of the vector. The operation of the root sign serves as an approximate standardization. Since after the pyramid pooling operation, the lengths of K and V are shorter than X, and Q only calculates attention with a portion of key–value pairs that have a strong correlation; thus, PDSA is more efficient than traditional MHSA. Furthermore, due to the highly abstract multiscale information contained in K and V, their ability in global context modeling is more prominent.

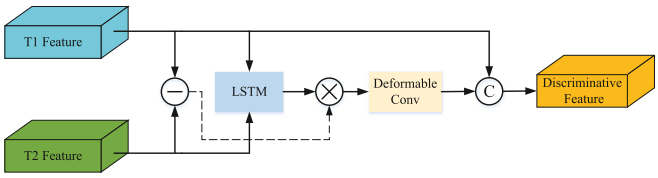

Fig. 4. Illustration of the DEM.

2) Global Semantic Enhancement Module: The designed GSEM refers to the standard structure of ViT, which is composed of a PDSA module and a multilayer perceptron (MLP) module, as shown in Fig. 3(b). First, the embedded token sequence of the input bitemporal images is layer normalized to ensure that the input values are not too large to be processed. Then, input it to the PDSA module and residual connect its output with the embedded sequence. Subsequently, another layer normalization is applied, and the feedforward network uses MLP for feature projection, followed by a residual connection. The previous calculation process can be defined as follows:

$$X_{\text{att}} = \text{PDSA}(LN(X)) + X \tag{10}$$

$$X_{\text{out}} = \text{MLP}(LN(X_{\text{att}})) + X_{\text{att}} \tag{11}$$

where X, X_{att}, and X_{out} denote the input, output of the PDSA module, and the encoded output of GSEM, respectively.

D. Difference Enhancement Module

Multiscale feature extraction enhances the feature expression ability of the model to a certain extent. However, during the procedure of feature encoding and decoding, it is easy to encounter the problem of high-frequency detail information loss, which leads to the blurring of boundaries between different levels of semantics and, thus, easily loses small target changes. To alleviate this issue, a DEM, consisting of two convolutional units and an SA module, is proposed by Li et al. [56] to provide weights for deep features. However, limited by the convolutional receptive field, its enhancement effect is limited.

To further refine the differences between bitemporal features and maximize the difference between changed semantics and unchanged semantics, we design a DEM through two stages, as shown in Fig. 4.

In the first stage, we enhance differential information through an adaptive weighting mechanism. Inspired by Bai et al. [17], we introduce long short-term memory (LSTM) to analyze the temporal correlation between input bitemporal images as the adaptive weighting adjustment. In addition, bitemporal images often exhibit spatial–spectral differences caused by variations in lighting and viewing angle. Linear differential analysis methods struggle to provide accurate CD results, whereas LSTM, as a nonlinear differential analysis method, effectively learns spectral–temporal feature representations between image pairs, thereby enhancing the difference information. When the distance between the original feature pairs is large, assign a higher weight. Since a larger distance implies a higher probability of being part of a changing region. On the contrary, a lower weight will be assigned. This procedure can be defined as follows:

$$D_i = |F_1 - F_2| \odot W_i \tag{12}$$

where F_1 and F_2 denote the bitemporal features of the input, respectively. W denotes the adaptive weight matrix. D denotes the enhanced feature bias.

In the second stage, we propose using deformable convolution to further expand the semantic differences between the changed regions and unchanged regions. Subsequently, feature fusion is completed through skip connection propagation to obtain enhanced features. Different positions in the image may correspond to objects with different scales or irregular shapes, and using traditional convolution inevitably has the drawback of fixed geometric structures. In particular, deformable convolution can change the range of the receptive field by introducing an offset ΔI_n to make the convolution kernel scalable. Taking the 3×3 convolution for instance, for 9 positions $B = \{(-1,-1),(-1,0),\ldots,(1,1)\}$ with an expansion coefficient of 1, output features in position I_0 can be defined as follows:

$$y(I_0) = \sum_{I_n} w(I_n) \cdot x(I_0 + I_n + \Delta I_n) \tag{13}$$

where I_0 denotes a point on the input feature map x. I_n denotes the offset of each point in the convolutional kernel relative to its center point. $w(I_n)$ denotes the weight at the corresponding position in the convolutional kernel. Our intuition is that deformable convolution can learn and match changed semantic boundaries with different directions and irregular shapes, maximizing the difference between changed areas and unchanged areas, thereby enhancing the discriminative capability of bitemporal image features.

E. Multiscale Fusion Decoder

The output by the GSEM is high-level semantic features, focusing on mining the global spatiotemporal relationships between objects. The multiscale low-level features extracted by EfficientNetV2-S contain rich local detail information, which can help better reconstruct the spatial structure of changed regions. Considering the symmetric structure of the encoder–decoder, the designed decoder of DPDANet consists of four upsampling modules. We adjust the number of channels through 3×3 convolution and use bicubic interpolation for upsampling. During the decoding process, multiscale feature aggregation is accomplished by merging the upsampled deep features with the resolution-matched shallow features through skip connections, resulting in more compact and integrated feature representations.

F. Loss Function

During the generation process of the change map, there may be a serious imbalance in the number of pixels between the changed regions and unchanged regions in the image. Traditional loss function processing may merge a few changed pixels into a much larger group of unchanged ones, resulting in the missed detection of small target changes. To alleviate the impact of unbalanced samples, we introduce an adaptive

weighted binary cross-entropy loss function and a weighted dice loss function [57].

The binary cross-entropy loss is often used in the binary classification problem. To alleviate the issue of unbalanced samples, the weight is introduced into the changed and unchanged categories to obtain an adaptive weighted binary cross-entropy loss function. Its definition is as follows:

$$L_{\text{wbce}} = -\frac{1}{N}\sum_{i=0}^{N} w_c t_i \log(p_i) + w_u(1-t_i)\log(1-p_i) \quad (14)$$

$$w_c = f_{\text{scale}}\left(\sum_{i=0}^{n_c}(1-p_i)/n_c\right) \quad (15)$$

$$w_u = f_{\text{scale}}\left(\sum_{i=0}^{n_u} p_i/n_u\right) \quad (16)$$

where t_i denotes the truth value of the pixel i, and p_i denotes the predicted value, with a value of 1 when a change occurs and 0 when no change occurs. Note that, weights are represented by pixel proportions. w_c and w_u denote the weight of the changed category and unchanged category, respectively. n_c and n_u denote the total number of changed pixels and unchanged pixels, respectively. f_{scale} denotes the proportional function. The adaptive weighted binary cross-entropy loss function can effectively avoid the tendency toward a certain category of the model, resulting in more robust results.

Dice loss is a region-based loss function commonly used in image segmentation tasks, which can alleviate the problem of sample imbalance. Its definition is as follows:

$$L_{\text{dice}} = 1 - \frac{2\times\sum_{i=0}^{N} t_i p_i}{\sum_{i=0}^{N}(t_i+p_i)+\varepsilon} \quad (17)$$

where ε denotes a constant (set to 1e-7 by default). We use the combination of the previous two loss functions as the optimization objective to enhance the training stability of small target changes and improve the detection accuracy. Then, the loss function in this article can be defined as follows:

$$L = L_{\text{wbce}} + \lambda L_{\text{dice}} \quad (18)$$

where λ denotes the weight used to balance L_{wbce} and L_{dice}, which is set to 1 based on experience in this article.

IV. Experiments

In this section, we first introduce the adopted datasets, then the implementation details of the proposed DPDANet, and the evaluation metrics are described briefly. Finally, we give the experimental results and the detailed analysis.

A. Dataset Descriptions

To evaluate the proposed DPDANet, we conduct extensive experiments on four public datasets: 1) CDNet-2014 [58], 2) LEVIR-CD [35], 3) SYSU-CD [59], and 4) CDD [60].

1) The CDNet-2014 dataset contains a total of 31 video sequences, captured by ordinary optical cameras and near-infrared cameras with different resolutions, covering different indoor and outdoor scenes. This dataset contains 11 categories, fully considering covariate factors, such as dynamic background, camera shaking, shadows, infrared, and lighting. We followed the same dataset split as described in [58] and extracted samples from different scenarios within each category. Specifically, the training, validation, and test sets images consisted of 73 276, 18 319, and 18 319 pairs of images, respectively. To facilitate comparisons with other algorithms, we normalize all image sizes to 256×256.
2) The LEVIR-CD dataset contains 637 pairs of remote-sensing images with a size of 1024×1024 and the spatial resolution of 0.5 m, covering covariance factors such as season and lighting changes. We followed the same dataset split as described in [35], yielding 445/64/128 pairs of images for the training, validation, and test sets, respectively. Considering hardware limitations, we cut the image into small patches of size 256×256 in a nonoverlapping manner. Thus, the training set, validation set, and test set are expanded to 7120/1024/2048 pairs of images, respectively.
3) The SYSU-CD dataset is a large-scale remote-sensing image CD dataset recently released by Sun Yat-sen University. It contains 20 000 pairs of images with a size of 256×256. We followed the same dataset split as described in [59], yielding 12 000/4000/4000 pairs of images for the training, validation, and test sets, respectively. This dataset provides various types of complex change scenes, including ships, roads, urban buildings, and changes in vegetation.
4) The CDD dataset is a real seasonal variation dataset captured by Google Earth, consisting of 7 pairs of 4725×2700 pixel images and 4 pairs of 1900×1000 pixel images with a spatial resolution of 3–100 cm/pixel. The dataset captures various changes caused by buildings, roads, cars, and other factors. To prepare the dataset, we followed the methodology described in [60]. We split all images into 256×256 image patches, and then 15 998 pairs of images were obtained through image enhancement methods. The number of the training set, validation set, and test set images is 10 000/2998/3000 pairs, respectively.

B. Implementation Details and Evaluation Metrics

We implement the proposed DPDANet using Python in conjunction with the PyTorch library and conduct all experiments on a single NVIDIA Tesla A100 GPU. We use random horizontal flipping, random vertical flipping, and random cropping to enhance the training data. We use EfficientNetV2-S as the backbone network and accelerate the training process through the pretrained weight. Due to the limitation of hardware, the batch size is set to 8. We randomly initialize the network and use the Adam optimizer with an initial learning rate of 1e-4, a momentum of 0.999, and a weight decay of 5e-4 for model

optimization. We train 200 epochs to achieve convergence of the model.

To accurately evaluate the performance of our proposed model, we use four common quantitative evaluation metrics: 1) precision (P), 2) recall (R), 3) F1-score, and 4) Intersection over Union (IoU). Their definitions are as follows:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

$$\text{F1-score} = \frac{2PR}{P + R} \tag{21}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{22}$$

where true positive (TP), false positive (FP), and false negative (FN) represent the number of unchanged pixels detected correctly, unchanged pixels unpredicted, and changed pixels unpredicted, respectively.

C. Comparative Experiments

1) Comparative Methods: To verify the performance of the proposed DPDANet, we select seven SOTA methods for experimental comparison, including spatiotemporal attention network (STANet) [59], BIT [45], multiscale twin parallel convolutional network (MSPSNet) [61], densely connected Siamese network (SNUNet) [29], improved separable deep network (ISNet) [62], Siamese U-shaped MLP-based network (SUMLP) [22], and W-shaped hierarchical network (WNet) [63].

a) *STANet:* STANet is a multiscale feature extraction network based on ResNet, which proposes spatiotemporal attention modules and pyramid spatiotemporal attention modules to refine feature representations.
b) *BIT:* BIT is a transformer-based network. It adopts a postfusion strategy and can achieve excellent performance with the outstanding long-range context modeling ability of the transformer.
c) *MSPSNet:* MSPSNet is a multiscale Siamese network based on a parallel convolutional structure, which uses channel attention to enhance the information representation of features.
d) *SNUNet:* SNUNet is a densely connected Siamese network used for very high resolution image CD. It suppresses localization errors to a certain extent by refining features at different semantic levels.
e) *ISNet:* ISNet is an improved separable deep learning network that extracts highly discriminative hierarchical features through a proposed boundary maximization strategy combined with channel and spatial attention.
f) *SUMLP:* SUMLP is a Siamese U-shaped CD network entirely based on MLP architecture. It adopts CycleMLP blocks as the basic units of the network and eliminates convolutional andSA operations, achieving excellent performance.
g) *WNet:* WNet is a W-shaped Siamese network for high-resolution remote-sensing image CD. It integrates a Siamese CNN and a Siamese transformer in the encoder, aiming to simultaneously model multiscale local details and global context relationships.

2) Qualitative Comparison: Figs. 5–8 show the visual qualitative comparison results of different methods on the CDNet-2014, LEVIR-CD, SYSU-CD, and CDD datasets, respectively. For better visualization, we adopt several colors to represent TP (white), TN (black), FN (red), and FP (green). Overall, profit by the proposed DEM to address boundary blur issues, the edge details in the change map produced by DPDANet are superior and exhibit a closer resemblance to the ground truth image.

In the perspective change scenario of the CDNet-2014 dataset, the accurate location of changed areas and effective suppression of perspective change interference are achieved by DPDANet, whereas STANet, BIT, MSPSNet, and SNUNet struggle to completely locate the changed areas, resulting in numerous false detections and missed detections, as demonstrated in the second row of Fig. 5. In scenes with insufficient lighting at night, DPDANet can maintain the integrity of the boundaries as much as possible and generate refined change maps, as shown in the fourth row of Fig. 5. In the scenario of small target CD, both SUMLP and WNet showed varying levels of missed detections, whereas DPDANet successfully detected the farthest appearing car, as shown in the sixth row of Fig. 5. This highlights the exceptional performance of the developed PDSA mechanism in effectively detecting subtle changes in small targets.

We selected samples of small-sized buildings from the LEVIR-CD dataset for visual comparison. In the presence of strong illumination variations, both WNet and our DPDANet demonstrated significantly lower false detections and missed detections, resulting in clearer boundaries between buildings, as shown in the fourth and sixth rows of Fig. 6.

In the complex scenes of the SYSU-CD dataset, BIT and SNUNet exhibited relatively inferior performance in detecting newly constructed buildings within forest areas, as shown in the second and fourth rows of Fig. 7. However, in the majority of cases, DPDANet outperformed its counterparts by successfully capturing more comprehensive regions of change. Our proposed model was able to differentiate pseudochanges occurring within road areas that bear resemblance to the appearance of buildings.

In the season-varying CDD dataset, STANet, BIT, and MSPSNet demonstrate the ability to detect relatively prominent changes. However, they exhibit limitations in capturing fine-grained changes, such as cars (fourth row in Fig. 8) and small roads (sixth row in Fig. 8), potentially due to their subpar feature integration capabilities. In contrast, our DPDANet exhibits excellent noise suppression performance, resulting in change maps with more compact interiors and well-defined boundaries that closely approximate the ground truth labels.

3) Quantitative Comparison: Tables I–IV present the quantitative comparison results of the proposed model and benchmark models in the CDNet-2014, LEVIR-CD, SYSU-CD, and CDD test sets, respectively. It can be found that quantitative comparison results confirm the intuitive visual comparison results. Our DPDANet outperformed other benchmark models in various metrics in CDNet-2014, SYSU-CD, and CDD datasets and

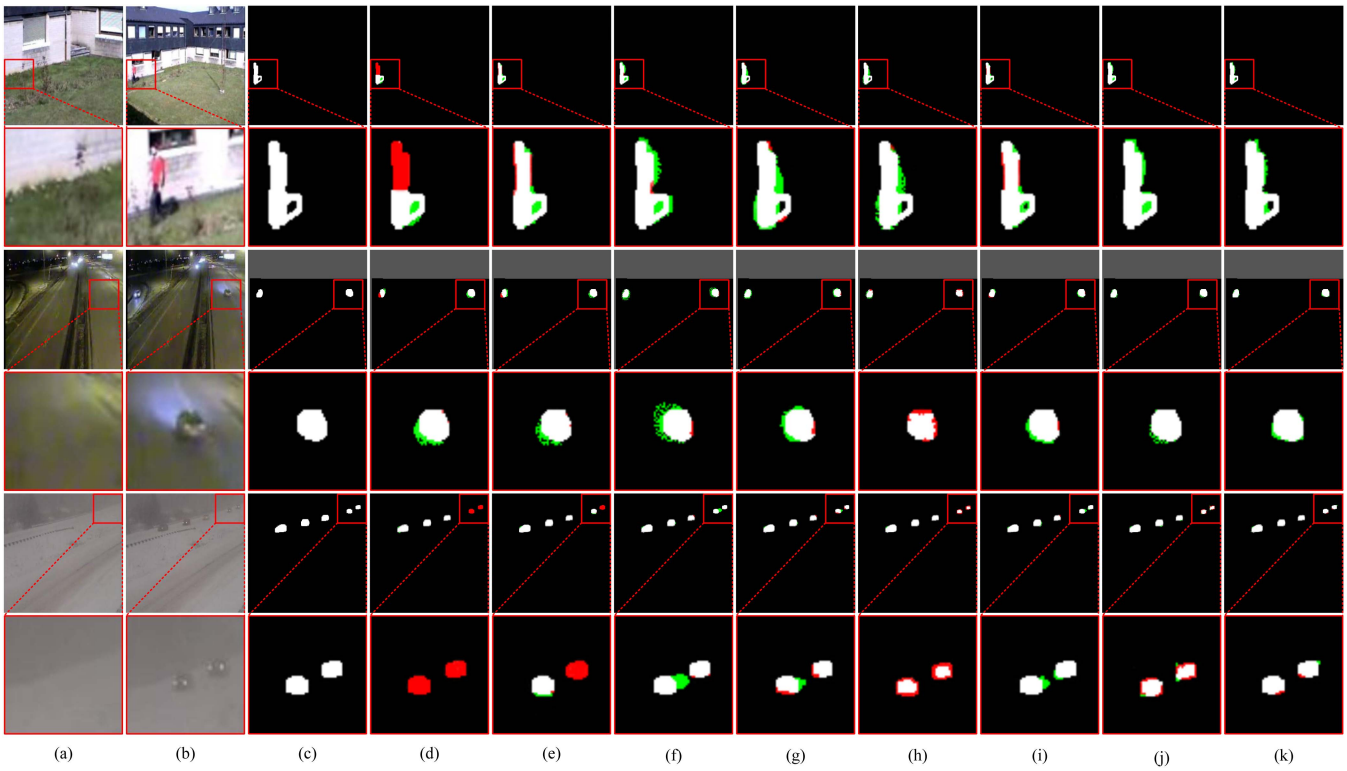

Fig. 5. Qualitative comparisons on CDNet-2014 dataset. (a) T_0 temporal image. (b) T_1 temporal image. (c) Ground truth. (d) STANet. (e) BIT. (f) MSPSNet. (g) SNUNet. (h) ISNet. (i) SUMLP. (j) WNet. (k) Proposed DPDANet. Color: white for TP (i.e., "changed"), black for TN (i.e., "unchanged"), red for FN, and green for FP.

TABLE I
QUANTITATIVE COMPARISON RESULTS ON CDNET-2014 DATASET

Model	P (%)	R (%)	F1 (%)	IoU (%)
STANet [59]	93.83	80.36	85.91	73.24
BIT [45]	93.60	88.97	90.45	74.67
MSPSNet [61]	96.95	85.32	91.18	81.57
SNUNet [29]	96.94	87.29	92.04	82.43
ISNet [62]	96.82	87.45	91.76	82.07
SUMLP [22]	96.75	87.37	91.92	81.88
WNet [63]	96.96	87.81	92.15	82.59
DPDANet (ours)	**97.89**	**91.13**	**93.05**	**84.79**

The bold entities denote the best results.

TABLE II
QUANTITATIVE COMPARISON RESULTS ON LEVIR-CD DATASET

Model	P (%)	R (%)	F1 (%)	IoU (%)
STANet [59]	83.81	91.03	87.35	78.46
BIT [45]	89.24	89.37	89.31	80.68
MSPSNet [61]	91.38	87.08	89.18	82.17
SNUNet [29]	91.80	88.53	90.14	82.04
ISNet [62]	92.46	88.27	90.32	82.35
SUMLP [22]	89.88	**91.65**	90.75	82.46
WNet [63]	91.16	90.18	90.67	82.93
DPDANet (ours)	**93.17**	88.98	**91.92**	**83.37**

The bold entities denote the best results.

TABLE III
QUANTITATIVE COMPARISON RESULTS ON SYSU-CD DATASET

Model	P (%)	R (%)	F1 (%)	IoU (%)
STANet [59]	70.28	82.88	75.21	61.50
BIT [45]	79.18	77.01	78.08	64.04
MSPSNet [61]	76.14	79.93	77.39	64.18
SNUNet [29]	78.41	73.38	75.81	61.04
ISNet [62]	80.27	76.41	78.29	64.44
SUMLP [22]	74.81	81.86	78.38	64.18
WNet [63]	81.71	79.58	80.64	67.55
DPDANet (ours)	**85.03**	**82.89**	**82.08**	**70.67**

The bold entities denote the best results.

TABLE IV
QUANTITATIVE COMPARISON RESULTS ON CDD DATASET

Model	P (%)	R (%)	F1 (%)	IoU (%)
STANet [59]	93.10	93.86	93.48	87.76
BIT [45]	95.31	87.31	91.13	83.71
MSPSNet [61]	95.83	93.49	94.65	89.36
SNUNet [29]	94.82	92.45	93.62	88.10
ISNet [62]	95.18	94.43	94.80	90.12
SUMLP [22]	98.63	96.82	97.72	91.18
WNet [63]	96.95	97.32	97.23	93.99
DPDANet (ours)	**98.75**	**97.89**	**97.73**	**94.27**

The bold entities denote the best results.

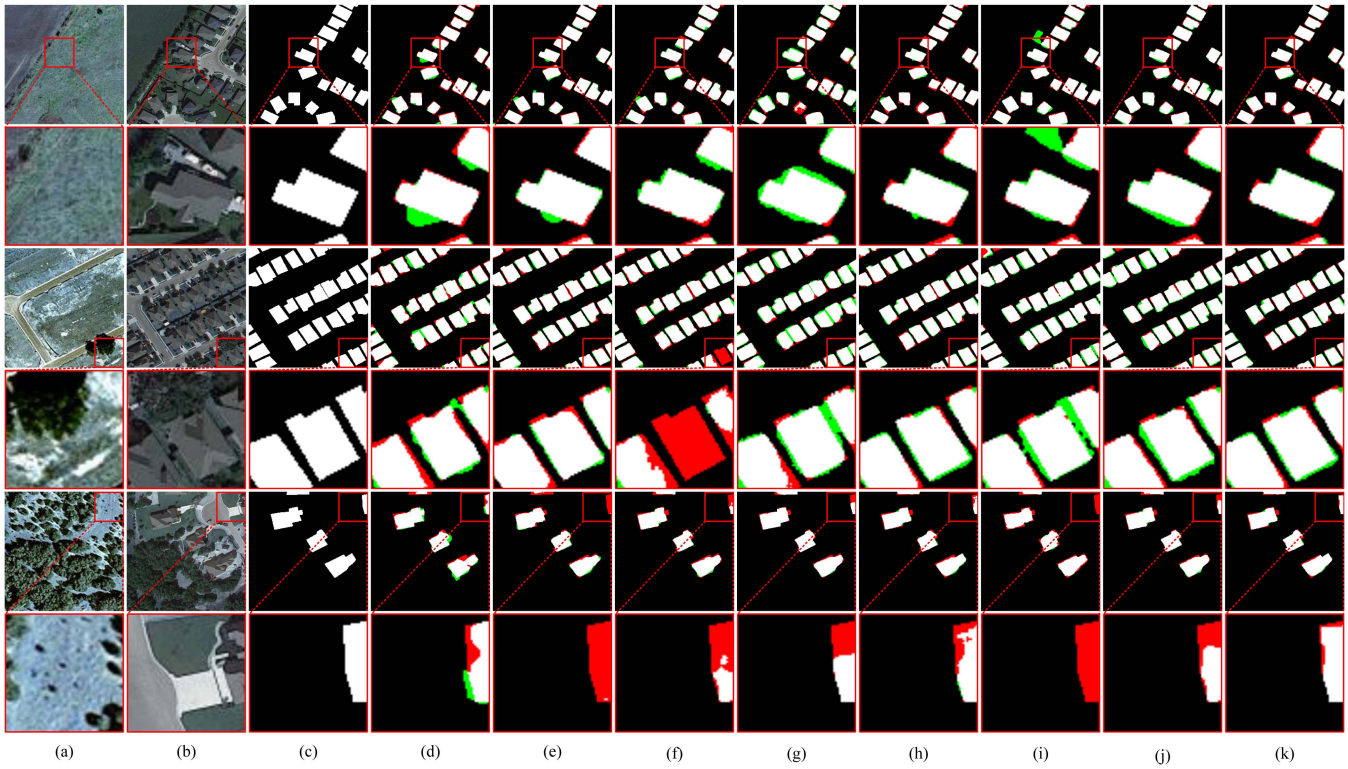

Fig. 6. Qualitative comparisons on LEVIR-CD dataset. (a) T_0 temporal image. (b) T_1 temporal image. (c) Ground truth. (d) STANet. (e) BIT. (f) MSPSNet. (g) SNUNet. (h) ISNet. (i) SUMLP. (j) WNet. (k) Proposed DPDANet. Color: white for TP (i.e., "changed"), black for TN (i.e., "unchanged"), red for FN, and green for FP.

achieved the best overall performance in the LEVIR-CD dataset, demonstrating the strong generalization capability of our model.

Specifically, in the CDNet-2014 dataset, DPDANet demonstrates outstanding performance with a P of 97.89%, R of 91.13%, F1-score of 93.05%, and IoU of 84.79%. Notably, these metrics outperform the second-best model (WNet) in the same category by margins of 0.93% in P, 3.32% in R, 0.9% in F1-score, and 2.2% in IoU. These results solidify the superiority of DPDANet in terms of quantitative evaluation compared to its closest competitor. In the LEVIR-CD dataset, the P and F1-score of WNet are 91.16% and 90.67%, respectively. This indicates that its multilevel local fine-grained features and global long-range contextual dependencies effectively enhance feature representation. While our DPDANet achieved optimal P, and F1-score. The positive outcomes can be attributed to three key factors. First, the utilization of both CNN and transformer in DPDANet effectively capitalizes on their respective advantages, facilitating comprehensive learning of local multiscale features and global semantic relationships. Second, the developed DEM successfully highlights the regions of change between bitemporal images, effectively reducing instances of missed detections and false detections and mitigating the issue of blurred boundaries. Third, the proposed PDSA mechanism offers an efficient approach to modeling global semantic relationships while minimizing the interference caused by noise variations. Collectively, these factors contribute to the generation of promising results.

TABLE V
PARAMETER COMPARISON OF DIFFERENT MODELS

Model	Param. (M)	FLOPs (G)	Inference Time (ms)
STANet [59]	16.93	**6.58**	97.12
BIT [45]	11.47	26.31	**33.96**
MSPSNet [61]	**2.21**	14.17	38.99
SNUNet [29]	3.01	27.56	34.23
ISNet [62]	18.01	11.45	68.35
SUMLP [22]	41.63	9.08	71.93
WNet [63]	43.07	19.20	57.69
DPDANet (ours)	32.17	13.41	43.38

The bold entities denote the best results.

4) Parameter Comparison: We compare the number of parameters (Params.), the computational complexity (floating-point operations, FLOPs), and the inference time on GPU between DPDANet and several benchmark models, as shown in Table V. Thanks to the token sequence length reduction and the sparse indexing operations enabled by the PDSA mechanism, DPDANet exhibits significantly reduced computational complexity compared to similar models such as BIT and WNet. In comparison to other models, our proposed model does not exhibit particularly noticeable advantages in terms of parameter size and inference time. This is primarily due to the high memory

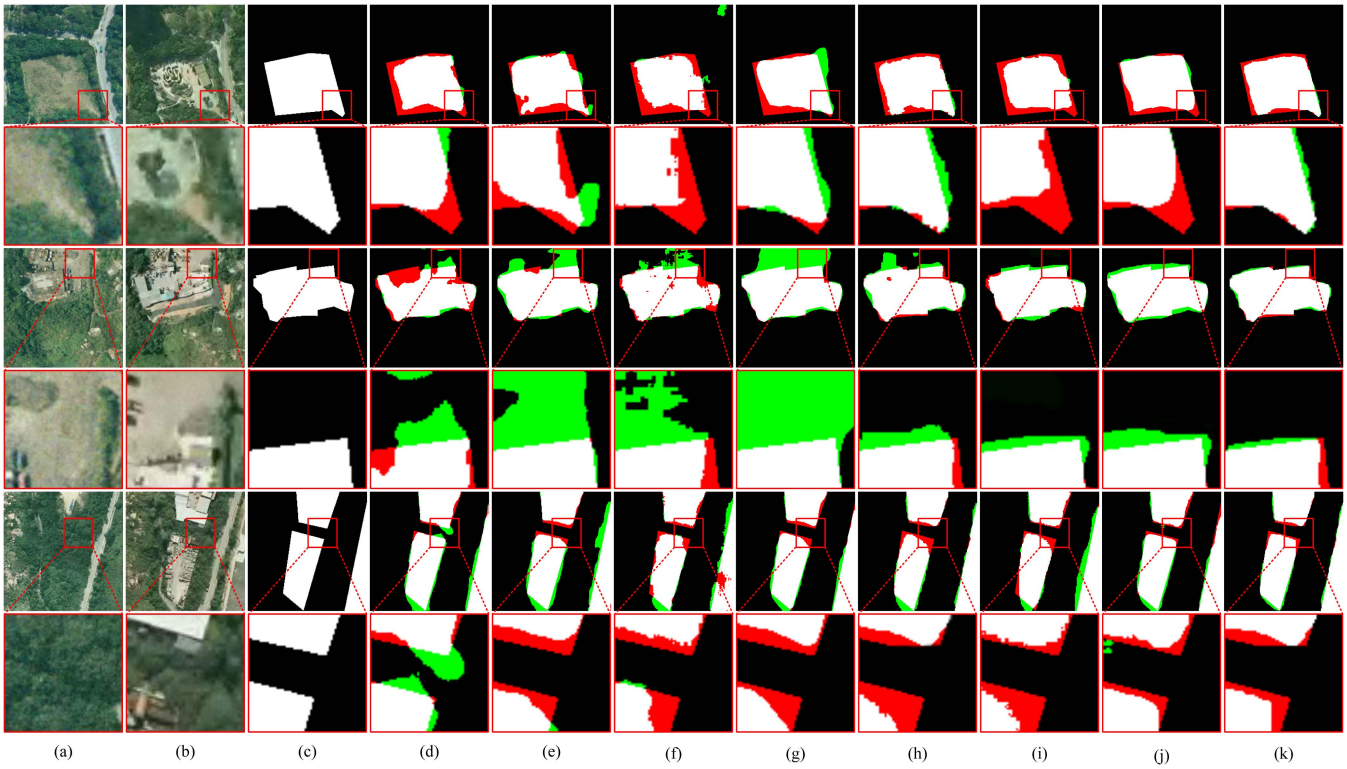

Fig. 7. Qualitative comparisons on SYSU-CD dataset. (a) T_0 temporal image. (b) T_1 temporal image. (c) Ground truth. (d) STANet. (e) BIT. (f) MSPSNet. (g) SNUNet. (h) ISNet. (i) SUMLP. (j) WNet. (k) Proposed DPDANet. Color: white for TP (i.e., "changed"), black for TN (i.e., "unchanged"), red for FN, and green for FP.

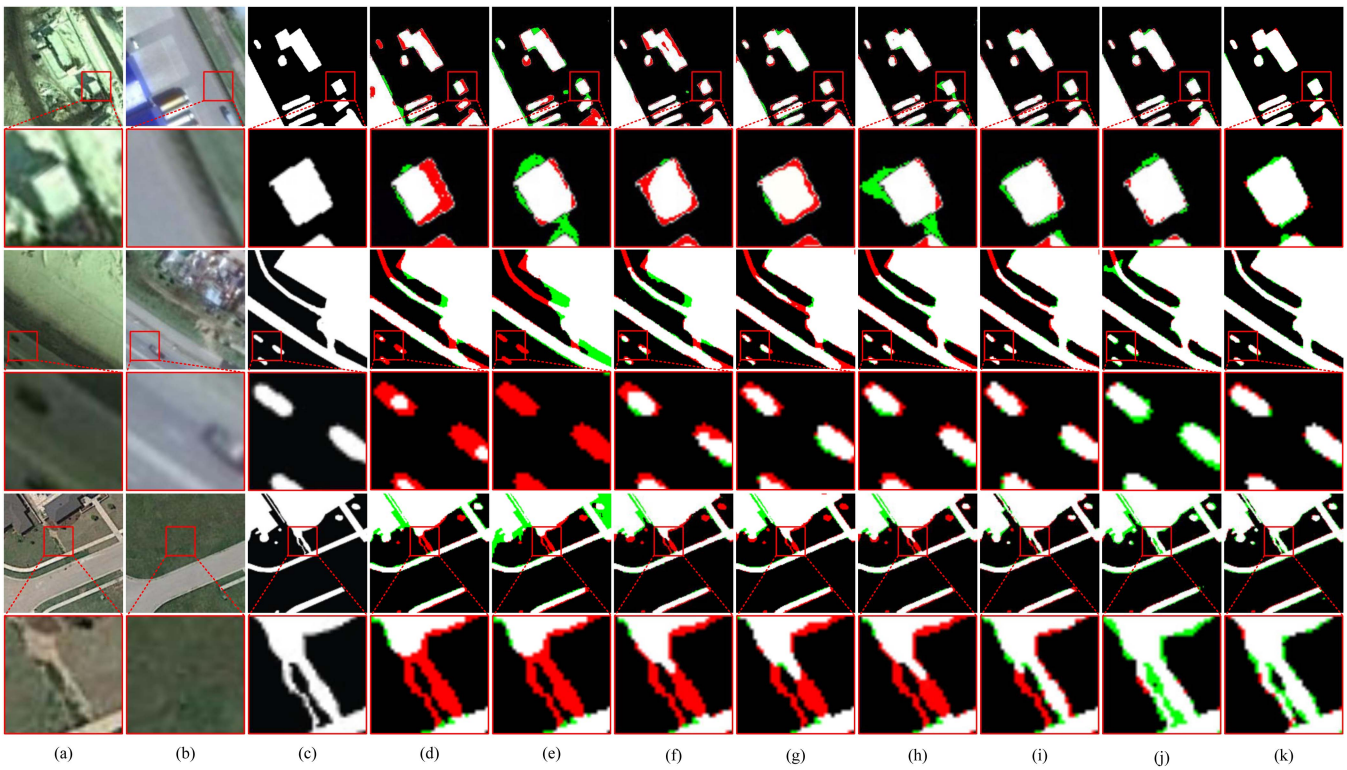

Fig. 8. Qualitative comparisons on CDD dataset. (a) T_0 temporal image. (b) T_1 temporal image. (c) Ground truth. (d) STANet. (e) BIT. (f) MSPSNet. (g) SNUNet. (h) ISNet. (i) SUMLP. (j) WNet. (k) Proposed DPDANet. Color: white for TP (i.e., "changed"), black for TN (i.e., "unchanged"), red for FN, and green for FP.

TABLE VI
ABLATION STUDY ON MODULE VERIFICATION

Method	P(%)	R(%)	F1 (%)	IoU (%)
w/o DEM & GSEM	94.07	90.02	91.12	82.04
w/o DEM	95.34	**91.34**	92.47	82.71
DEM [56]	96.28	91.14	92.49	83.21
w/o GSEM	95.16	91.09	91.27	82.32
SA	96.97	91.04	92.18	83.59
MSA [25]	97.03	91.11	92.39	83.88
DPDANet (ours)	**97.89**	91.13	**93.05**	**84.79**

The bold entities denote the best results.

TABLE VII
INFLUENCE OF DIFFERENT BACKBONES

Backbones	Param. (M)	FLOPs (G)	P (%)	R (%)	F1 (%)
EfficientNetV2-S	32.17	**13.41**	97.89	91.13	93.05
ResNet18	**19.50**	18.42	98.06	90.74	92.51
ResNet50	33.34	44.96	97.84	91.06	92.88
ResNet101	56.68	63.27	**98.12**	**91.67**	**93.21**

The bold entities denote the best results.

access overhead associated with the depthwise convolutions employed by the EfficientNetV2-S backbone network. Factors such as memory access and memory occupation play a crucial role in influencing the inference time of the model. In light of the heightened accuracy of our model, the effect of increased parameters can be regarded as minimal. In comparison to the similar model WNet, the proposed DPDANet showcases a notable reduction of 25.31% in parameter size. Moreover, it achieves a commendable decrease of 24.80% in inference time on GPU. These results effectively confirm the superior performance of DPDANet. In sum, our model's parameter size, FLOPs, and inference time stay in an acceptable range.

D. Ablation Study

In this section, we conduct extensive ablation experiments on the DPDANet model on three datasets.

1) Module Verification Experiment: We conducted ablation experiments on the CDNet-2014 dataset to validate the effectiveness of each component in DPDANet. By systematically removing or replacing different modules, we assessed their contributions to the overall performance. The experimental results are summarized in Table VI, where "w/o" is the abbreviation for "without," indicating the removal of a module from the network. Specifically, "DEM" refers to our proposed DEM while "DEM [56]" refers to the DEM proposed in [56]. "GSEM" represents the global semantic enhancement module, "SA" denotes the traditional SA module, and "MSA [25]" represents the moving-window SA module.

a) Effectiveness of DEM: Our motivation in designing the DEM was to refine the difference between bitemporal features, maximizing the dissimilarity between changed and unchanged semantics and enhancing the discriminability of changed regions. To evaluate the effectiveness of our proposed DEM, we conducted experiments by initially omitting this module and directly passing the concatenated bitemporal features to the corresponding layers of the decoder. The quantitative comparison results are presented in Table VI. The results demonstrably indicate a significant decline in performance when the DEM is removed. Moreover, we substituted our DEM with the DEM introduced in [56]. The results reveal that the P, F1-score, and IoU exhibited decreases of 1.61%, 0.56%, and 1.58%, respectively. These experimental findings substantiate the feasibility and superiority of our DEM.

b) Effectiveness of GSEM: In the design of the GSEM, we propose a PDSA mechanism, which efficiently captures multiscale contextual information by leveraging dynamic sparsity. To assess the performance enhancement provided by GSEM on the overall network, we, respectively, employed ordinary SA (SA) and moving-window SA (MSA) as substitutes for PDSA in constructing the GSEM. As presented in Table VI, removing the GSEM led to a significant performance decline. The inclusion of SA and MSA contributed to promising performance improvements. Nevertheless, GSEM with PDSA consistently outperformed SA and MSA across all metrics. These findings demonstrate that GSEM, based on the proposed PDSA mechanism, excels in modeling global contextual relationships and exhibits stronger robustness to environmental noise variations.

c) Influence of Different Backbones: We conducted additional experiments to verify the influence of different backbone networks on our model. Specifically, we employed ResNet18, ResNet50, and ResNet101 as substitutions for the default EfficientNetV2-S. From the results in Table VII, it was observed that DPDANet with EfficientNetV2-S outperformed DPDANet with ResNet18, showed comparable performance to DPDANet with ResNet50, and slightly underperformed compared to ResNet101. It is evident that deeper backbone networks generally offer detection performance gains but also come with a significant increase in computational burden. The results indicate that EfficientNetV2-S effectively serves as a decent backbone network, delivering satisfactory performance while reducing the computational load.

2) Parameter Verification Experiment: We validated the impact of training epoch number, patch size, and depth of GSEM on the performance of DPDANet on three datasets.

a) Effect of Training Epoch Number: Fig. 9 shows the detection results of DPDANet on the LEVIR-CD and SYSU-CD validation sets. As can be seen from these figures, the model can achieve good validation results in a short period of time and achieve convergence when reaching a certain epoch. Continuing to increase the number of epochs for training will not significantly improve the performance of the model. Therefore, in our experiments, we uniformly set the epoch number to 200 and attenuate the learning rate after 50 epochs.

b) Effect of the Patch Size: In the CDNet-2014 dataset, we set the patch sizes to 8, 16, and 32, respectively, to study their impact on CD accuracy, as shown in Table VIII. The experimental results indicate that the detection performance of the model is positively correlated with patch size. The probable reason is that the patch size determines the length of the

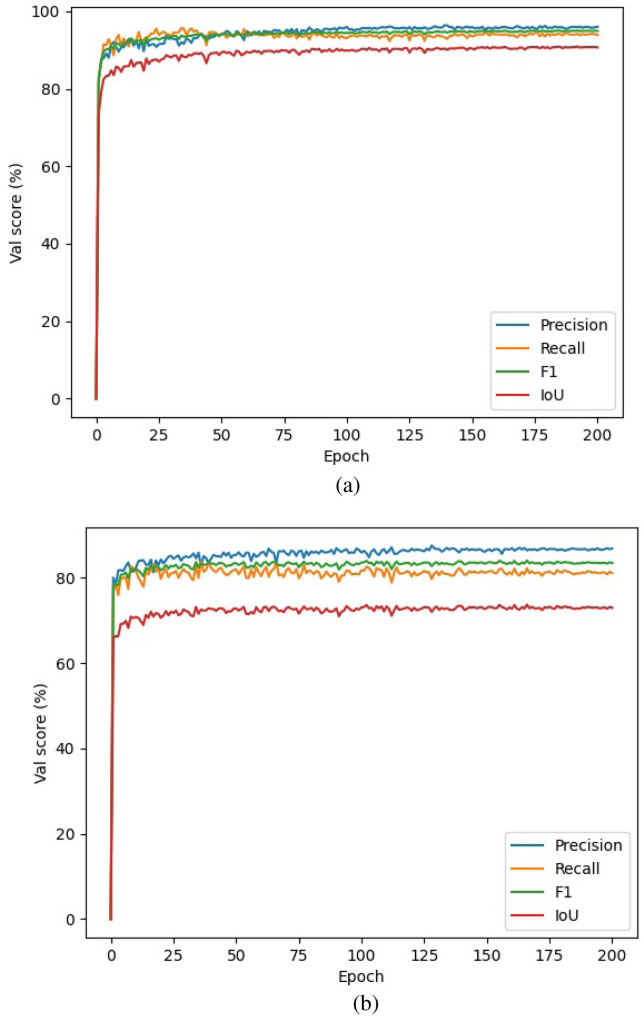

Fig. 9. Effect of training epoch number on (a) LEVIR-CD validation set and (b) SYSU-CD validation set.

TABLE VIII
EFFECT OF THE PATCH SIZE

Patch Size	*P* (%)	*R* (%)	F1 (%)	IoU (%)
8	97.71	91.03	93.01	84.69
16	97.89	91.13	93.05	84.79
32	**97.96**	**91.30**	**93.07**	**84.82**

The bold entities denote the best results.

generated token sequences. The longer the token sequence, the richer feature information it contains. However, considering the training efficiency, we set the patch size to 16 to balance the computational complexity and performance of the model.

c) Effect of the GSEM Depth: As shown in Table IX, the increase in GSEM depth slightly improves the detection performance of the model, but meanwhile brings about a rapid increase in parameter and computational complexity. Accordingly, in our experiments, the GSEM depth was set to 6 to balance efficiency and accuracy.

TABLE IX
EFFECT OF THE GSEM DEPTH

GSEM Depth	*P* (%)	*R* (%)	F1 (%)	Params (M)
1	97.63	91.08	92.88	**29.02**
6	97.89	91.13	93.05	32.17
12	**97.99**	**91.44**	**93.08**	34.25

The bold entities denote the best results.

E. Feature Visualization Experiment

We extracted a pair of images from the LEVIR-CD test set as an example, and Fig. 10 shows the feature visualization results of the 160th channel of four important nodes. Specifically, these four nodes refer to the concatenation and fusion of the outputs of the multiscale DEM and the GSEM. For the convenience of comparison, we normalized the feature maps of different nodes to the same size. Red indicates giving higher attention to potentially changed areas, whereas black indicates giving lower attention to unchanged areas.

It is evident that the proposed model has been learning the representation of changes at different stages. As the network depth increases, multiscale deep features are extracted progressively. The DEM and GSEM modules exhibit strong capability in discerning between changed and unchanged areas within the input feature map by combining global semantic information and local features. This enhances the expression of edges in changed areas and improves the performance of small target detection.

V. DISCUSSION

Our proposed DPDANet has demonstrated superior performance compared to existing comparative models through extensive experimental validation. As proposed, the EfficientNetV2-S backbone network enriches local detailed information through multiscale feature extraction and compensates for the deficiency in modeling long-range contextual information. By maximizing the semantic difference between changing and invariant regions, DEM effectively mitigates edge-blurring issues. In addition, the PDSA module, which combines pyramid pooling and dynamic sparse encoding, reduces the computational complexity of SA and captures more comprehensive long-range contextual information, thus enhancing the model's detection performance under complex environmental noise conditions.

Despite these achievements, our proposed method still has certain limitations, predominantly in terms of model parameter size and inference time. Due to the utilization of depthwise convolution as the basic units in the EfficientNetV2-S backbone network, it exhibits a high memory access overhead, leading to considerably slower inference times compared to regular convolutions. These factors have a notable impact on the real-time performance and operational efficiency of deploying the model on edge devices. Our future research will pursue two main directions. First, we will further investigate the development of efficient and lightweight backbone networks for enhanced feature extraction. Second, we will propose novel loss functions to address the issue of edge blurring from the perspective of optimizing the loss function.

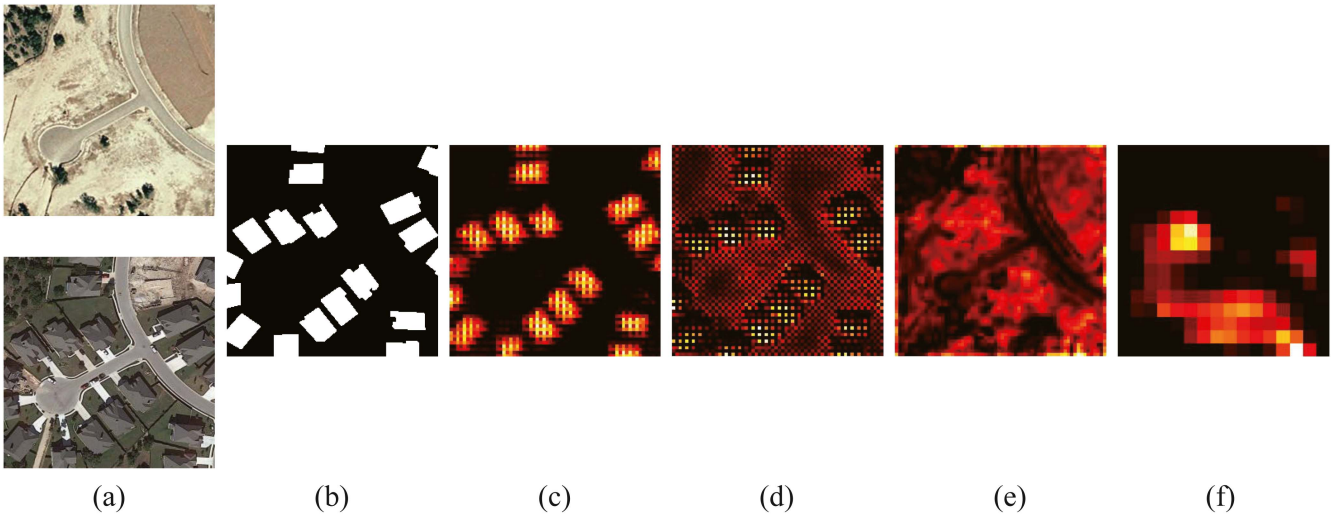

Fig. 10. Feature visualization. (a) Bitemporal images. (b) Ground truth. (c), (d), (e), and (f) 160th feature map of the first, second, third, and fourth scale outputs, respectively.

VI. Conclusion

In this article, we propose a new remote-sensing image CD network based on difference-enhanced pyramid pooling dynamic sparse attention (i.e., DPDANet). It is committed to solving multiscale and subtle CD, as well as blurry boundary issues. DPDANet utilizes EfficientNetV2-S as the baseline to extract multiscale local features, and integrates two proposed modules, DEM and GSEM, to improve its capability of detecting subtle changes and addressing blurry boundary issues. DEM obtains adaptive weight by introducing LSTM to analyze the temporal correlation of bitemporal images and combines deformable convolution to enhance the differential features of different semantics. GSEM simplifies SA computation by employing pyramid pooling operations and dynamic sparse encoding, reducing computational complexity while enhancing the modeling of global semantic relationships. We introduce a hybrid loss to complete the training process. Our DPDANet achieves SOTA performance on four public CD datasets, showcasing strong generalization ability and robustness against complex environments.

References

[1] G. I. Drakonakis, G. Tsagkatakis, K. Fotiadou, and P. Tsakalides, "OmbriaNet-supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2341–2356, Mar. 2022.

[2] G. Lei, A. Li, J. Bian, A. Naboureh, Z. Zhang, and X. Nan, "A simple and automatic method for detecting large-scale land cover changes without training data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7276–7292, Jul. 2023.

[3] J. Liu, J. Xiang, Y. Jin, R. Liu, J. Yan, and L. Wang, "Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: A survey," *Remote Sens.*, vol. 13, no. 21, Nov. 2021.

[4] W. Yuan, W. Ran, X. Shi, and R. Shibasaki, "Multiconstraint transformer-based automatic building extraction from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9164–9174, Sep. 2023.

[5] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, Sep. 2022.

[6] Y. Sun, K. Deng, K. Ren, J. Liu, C. Deng, and Y. Jin, "Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review," *ISPRS J. Photogramm. Remote Sens.*, vol. 208, pp. 14–38, Feb. 2024.

[7] A. C. Mondini, F. Guzzetti, P. Reichenbach, M. Rossi, M. Cardinali, and F. Ardizzone, "Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images," *Remote Sens. Environ.*, vol. 115, no. 7, pp. 1743–1757, Jul. 2011.

[8] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using TRI-temporal logic-verified change vector analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 278–293, Mar. 2020.

[9] D. K. S. Datta, "Colour band fusion and region enhancement of spectral image using multivariate histogram," *Int. J. Image Data Fusion*, vol. 12, no. 1, pp. 64–82, Mar. 2021.

[10] R. V. Fonseca, R. G. Negri, A. Pinheiro, and A. M. Atto, "Wavelet spatio-temporal change detection on multitemporal SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4013–4023, Apr. 2023.

[11] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5501212.

[12] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5526914.

[13] J. Wang, W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Hyperspectral and SAR image classification via multiscale interactive fusion network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10823–10837, Dec. 2023.

[14] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote sensing scene classification via multi-stage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5615312.

[15] J. Wang, W. Li, M. Zhang, and J. Chanussot, "Large kernel sparse ConvNet weighted by multi-frequency attention for remote sensing scene understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5626112.

[16] F. Zhou, C. Xu, R. Hang, R. Zhang, and Q. Liu, "Mining joint intra- and inter-image context for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 4403712.

[17] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5610613.

[18] Z. Li et al., "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5602812.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, 2015.

[21] 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.

[22] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SUMLP: A Siamese U-shaped MLP-based network for change detection," *Appl. Soft Comput.*, vol. 131, 2022.

[23] S. Chen, E. Xie, C. Ge, R. Chen, D. Liang, and P. Luo, "CycleMLP: A MLP-like architecture for dense visual predictions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14284–14300, Dec. 2023.

[24] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5224713.

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

[26] Z. Tu et al., "MaxViT: Multi-axis vision transformer," in *Computer Vision – ECCV*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham, Switzerland: Springer, Nov. 2022, pp. 459–479.

[27] M. H. Kesikoglu, U. H. Atasever, and C. Ozkan, "CrossFormer: A versatile vision transformer hinging on cross-scale attention," in *Proc. Int. Conf. Comput. Vis.*, 2022.

[28] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10323–10333.

[29] 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, Feb. 2022, Art. no. 8007805.

[30] M. Tan and Q. V Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 7102–7110.

[31] H. Du et al., "Bilateral semantic fusion Siamese network for change detection from multitemporal optical remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jun. 2022, Art. no. 6003405.

[32] 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.

[33] F. Y. Koltun, "Multi-scale context aggregation by dilated convolutions," *Comput. Sci.*, 2015.

[34] 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.

[35] 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.

[36] J. Chen et al., "DasNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, Nov. 2021.

[37] 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, Feb. 2021, Art. no. 8007805.

[38] 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.

[39] H. Chen, Z. Qi, and Z. Shi, "Efficient transformer based method for remote sensing image change detection," *J. Photogrammetry Remote Sens.*, 2021.

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

[41] F.-E. Jannat and A. R. Willis, "Improving classification of remotely sensed images with the Swin transformer," in *Proc. SoutheastCon*, 2022, pp. 611–618.

[42] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, vol. 139, pp. 7358–7367.

[43] J. Gu et al., "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12084–12093.

[44] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[45] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607514.

[46] 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, Apr. 2022, Art. no. 4410213.

[47] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[48] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6804–6815.

[49] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[50] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[51] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[52] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5617712.

[53] J. Yuan, L. Wang, and S. Cheng, "STransUNet: A Siamese TransUNet-based remote sensing image change detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9241–9253, Oct. 2022.

[54] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," p. 9, 2017, *arXiv:1704.04861*.

[55] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *Statistics*, 2016, *arXiv:1607.06450*.

[56] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5622519.

[57] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[58] W. Yi, P. M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "CDNet 2014: An expanded change detection benchmark dataset," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 387–394.

[59] 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, Jun. 2022, Art. no. 5604816.

[60] 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. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, Jun. 2018.

[61] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, "Deep multi-scale Siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5406512.

[62] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5623811.

[63] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "WNet: W-shaped hierarchical network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5615814.

Zhong Li received the B.E. degree in mechanical and electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2017, the M.E. degree in instrument and meter engineering from the National University of Defense Technology, Changsha, China, in 2019. He is currently working toward the Ph.D. degree in electrical engineering with the National Key Laboratory of Electromagnetic Energy, Naval University of Engineering, Wuhan.

His research interests include image processing, computer vision, and electrical intelligence and monitoring management technology.

Bin Ouyang received the B.E. degree in power system and automation, the M.S. degree in motor and electrical appliance, and the Ph.D. degree in electrical engineering from the Naval University of Engineering, Wuhan, China, in 2003, 2005, and 2011, respectively.

He is currently a Professor with the National Key Laboratory of Electromagnetic Energy, Naval University of Engineering. His research interests include computer vision and modeling, analysis, and control of electrical machines.

Shaohua Qiu received the B.E. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2011, and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2013 and 2018, respectively.

He is currently an Associate Professor with the National Key Laboratory of Electromagnetic Energy, Naval University of Engineering, Wuhan. His research interests include deep learning and image analysis.

Xinghua Xu received the B.S. and M.S. degrees in computer science and technology from the National University of Defense Technology, Changsha, China, in 2004 and 2007, respectively, and the Ph.D. degree in electrical engineering from the Naval University of Engineering, Wuhan, China, in 2017.

He is currently a Professor with the National Key Laboratory of Electromagnetic Energy, Naval University of Engineering. His research interests include deep learning, electromagnetic launch, and fault diagnosis.

Xiaopeng Cui received the B.E. degree in navigation engineering from Air Force No.1 Aviation University, Xinyang, China, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from the Naval University of Engineering, Wuhan, China, in 2009 and 2013, respectively.

He is currently an Associate Professor with the National Key Laboratory of Electromagnetic Energy, Naval University of Defense Technology. His research interests include electromagnetic launch and fault diagnosis.

Xia Hua received the B.S. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2016, and the M.S. and Ph.D. degrees in mechanical engineering from the Army Engineering University of PLA, Nanjing, China, in 2018 and 2021, respectively.

He is currently a Lecturer with the National Key Laboratory of Electromagnetic Energy, Naval University of Defense Technology, Wuhan, China. His research interests include computer vision, video compression, pattern recognition, and artificial intelligence.