This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/ICASSP49660.2025.10890539, ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)Demo

Dynamic Sparse Encoding and Cross-Temporal Attention for Remote Sensing Image Change Detection

Shaoxiong Lin
Shaanxi Joint Laboratory of Artificial Intelligence Shaanxi University of Science and Technology
Xi'an, China
231611039@sust.edu.cn

Tao Lei ✉
Shaanxi Joint Laboratory of Artificial Intelligence Shaanxi University of Science and Technology
Xi'an, China
leitaoly@163.com

Tongfei Liu
Shaanxi Joint Laboratory of Artificial Intelligence Shaanxi University of Science and Technology
Xi'an, China
liutongfei_home@hotmail.com

Shuxin Zhang
Shaanxi Joint Laboratory of Artificial Intelligence Shaanxi University of Science and Technology
Xi'an, China
894662257@qq.com

Chongdan Min
Shaanxi Joint Laboratory of Artificial Intelligence Shaanxi University of Science and Technology
Xi'an, China
bs221611001@sust.edu.cn

Asoke K. Nandi
Department of Electronic and Electrical Engineering Brunel University London
London, United Kindom
asoke.nandi@brunel.ac.uk

***Abstract*—Due to the inherent inductive bias of operations, convolutional neural networks (CNN) cannot model global information of remote sensing (RS) images. In contrast, Transformer-based methods can establish long-range dependencies of images through self-attention (SA) mechanism, but it faces the challenges of computational complexity and memory requirements, but also ignores the exploration on the feature redundancy removal of RS images. To address these two issues, we propose a network based on dynamic sparse encoding and cross-temporal collaborative attention (DSECTCA-Net) for RS image change detection (CD). First, we implement dynamic sparse encoding (DSE) by designing hierarchical sparse Transformer module (HSTM), which decreases the correlation calculation of the SA mechanism and effectively reduces the computational complexity and parameter amount of Transformer. Secondly, we propose cross-temporal collaborative attention (CTCA) to model RS images in time series and fully explore the interactivity between dual-temporal RS images, so as to better extract the global understanding of visual scenes. Extensive experiments on two large-scale public RS datasets show that the proposed method not only provides higher detection accuracy, but also achieves lower computational complexity and required storage space than most popular CD networks.**

***Index Terms*—remote sensing image, change detection, sparse encoding, collaborative attention, Transformer.**

I. INTRODUCTION

Change detection (CD) is a technique that used to identify and analyze changes of surface features over time by utilizing remote sensing (RS) images and related geospatial data of the same area at different times. It has been widely applied in many fields such as urban planning, farmland management and environmental monitoring [1], [2].

In recent years, the rapid development of deep learning has provided a broader exploration space for the research on CD tasks. Convolutional neural network (CNN) -based and Transformer-based methods have made significant progress. Although CNN-based methods can effectively learn local information of RS images, more RS images are required to compensate for their shortcomings due to the inherent inductive bias of convolution operations. In contrast, Transformer-based methods successfully capture the global information of images through the self-attention (SA) mechanism. However, these methods often face the challenges of a high computational complexity and a large memory requirement due to the calculation of the pairwise sequence correlations between all spatial positions [3]. To address this problem, many methods have been proposed [4], [5] to improve the performance in processing long sequences of RS images. These methods include reducing the complexity from quadratic to linear by changing the order of matrix multiplication or reducing computation by reducing the dimensionality of attention weights [6], such as local windows, hole windows and axial attention [7]. However, these sparse attention mechanisms usually rely on manually selected sparsity patterns, without fully exploring the correlation between feature vectors *query* (**Q**), *key* (**K**) and *value* (**V**). To address these issues mentioned above, we mainly made the following three contributions:

1) A dynamic sparse encoding (DSE) module is proposed

✉ Corresponding author: Tao Lei.

to focus on a small number of related sequences in a query-adaptive manner, thereby reducing the computation of irrelevant sequences and the feature redundancy.

2) A cross-temporal collaborative attention (CTCA) module is designed to fully explores the interaction feasibility between dual-temporal RS images through incorporating temporal information.
3) We propose a network based on dynamic sparse encoding and cross-temporal collaborative attention (DSECTCA-Net). Experimental results show that compared with existing popular methods, our method not only provides higher detection accuracy, but also requires less storage space and computational cost.

II. Method

A. Overall Network Structure

The structure of the proposed method DSECTCA-Net as shown in Fig. 1, mainly including the DSE module, the CTCA module, and the feature decoding module.

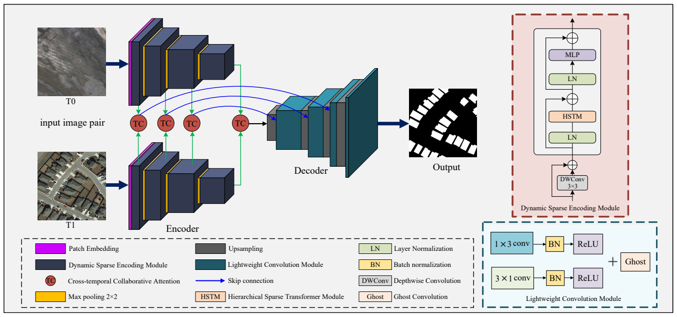
T0
input image pair
T1
TC
TC
TC
TC
Encoder
Decoder
Output
MLP
LN
HSTM
LN
DWConv 3×3
Dynamic Sparse Encoding Module
Patch Embedding
Dynamic Sparse Encoding Module
Cross-temporal Collaborative Attention
Max pooling 2×2
Upsampling
Lightweight Convolution Module
Skip connection
HSTM Hierarchical Sparse Transformer Module
LN Layer Normalization
BN Batch normalization
DWConv Depthwise Convolution
Ghost Ghost Convolution
1×3 conv
BN
ReLU
3×1 conv
BN
ReLU
+
Ghost
Lightweight Convolution Module

Fig. 1. The overall structure of DSECTCA-Net.

In the coding phase, the RS images are firstly segmented into patches, where each patch's feature dimensions are projected into arbitrary dimensions via patch embedding. Second, multiple DSE modules and downsampling layers are utilized to extract different feature representations with global information, aiming to optimize the model performance while reducing the computational cost. Finally, at each stage of the encoder, the feature maps obtained from the corresponding layers of the dual branches are fed into the CTCA module. These refined feature maps are subsequently skip-connected to the corresponding layers in the decoder, enriching change target feature maps.

In the decoding phase, the decoder consists of upsampling layers and lightweight convolution modules. To improve contour detection and robustness to image transformations (e.g., rotations or flips), we introduce the asymmetric ghost convolution within the lightweight convolution module [2], which can enhance the capture of horizontal and vertical edge information and reduce network parameters.

B. Dynamic Sparse Encoding Module

Unlike previous works [8], [9], the DSE module is specifically designed to efficiently capture the detailed features and long-range dependencies of RS images through the advanced visual Transformer. Specifically, a 3×3 depthwise convolution is utilized in the initial stage of the DSE module to capture the key position information. The feature maps are then normalized through a LayerNorm (LN) layer. Subsequently, a hierarchical sparse Transformer module (HSTM) is constructed to compute attention in a coarse-to-fine manner. Finally, the feature maps are processed through a multilayer perceptron (MLP) layer.

In the HSTM structure, the input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, is firstly divided into non-overlapping patches of size $S \times S$ through the patch embedding, and each patch contains HW/S^2 feature vectors. These feature vectors are linearly mapped to obtain $\mathbf{X} \in \mathbb{R}^{S^2 \times HW/S^2 \times C}$, and then subjected to three different linear transformations to obtain $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{S^2 \times HW/S^2 \times C}$ respectively. As shown in Fig. 2.

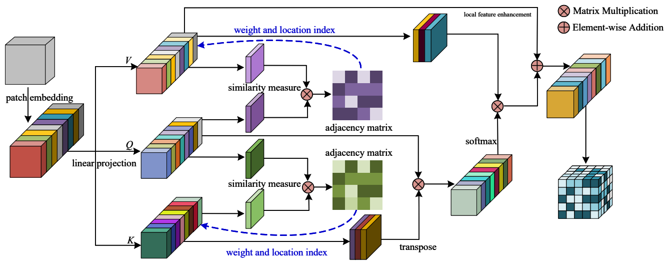

Fig. 2. The structure of HSTM.

Spatial averaging is performed on the obtained vectors $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ in each region to obtain vectors $\mathbf{Q}^{\mathrm{m}}$, $\mathbf{K}^{\mathrm{m}}$, $\mathbf{V}^{\mathrm{m}} \in \mathbb{R}^{S^2 \times C}$ separately. Then $\mathbf{Q}^{\mathrm{m}}$ is multiplied by the transpose of $\mathbf{K}^{\mathrm{m}}$ and $\mathbf{V}^{\mathrm{m}}$ respectively to construct region-to-region adjacency matrices $\mathbf{K}^{\mathrm{r}}$, $\mathbf{V}^{\mathrm{r}} \in \mathbb{R}^{S^2 \times S^2}$. Based on this, the top h most relevant regions are recorded. Their corresponding weights and indexes are saved in the weight index matrix. The top h most relevant regions $\mathbf{K}^h$ and $\mathbf{V}^h$, are selected and subjected to fine-grained matrix multiplication, followed by a non-linear activation function $\mathrm{Softmax}(\cdot)$ to obtain the attention matrix $\mathbf{A}_{\mathrm{att}}$. To supplement the local contextual information, we use the function $\mathrm{LE}(\cdot)$ to augment $\mathbf{V}$ with local information, employing depthwise convolution. The finally output is obtained as follows:

$$\mathbf{O} = \mathbf{A}_{\mathrm{att}} + \mathrm{LE}(\mathbf{V}) \tag{1}$$

Specifically, the detailed process of the DSE module is as follows:

$$\hat{\mathbf{C}}_{\mathrm{out}}^{l-1} = \mathrm{DW}(\mathbf{C}_{\mathrm{out}}^{l-1}) + \mathbf{C}_{\mathrm{out}}^{l-1} \tag{2}$$

$$\hat{\mathbf{C}}_{\mathrm{out}}^{l} = \mathrm{Attention}(\mathrm{LN}(\hat{\mathbf{C}}_{\mathrm{out}}^{l-1})) + \hat{\mathbf{C}}_{\mathrm{out}}^{l-1} \tag{3}$$

$$\mathbf{C}_{\mathrm{out}}^{l} = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{C}}_{\mathrm{out}}^{l})) + \hat{\mathbf{C}}_{\mathrm{out}}^{l} \tag{4}$$

where $\hat{\mathbf{C}}_{\mathrm{out}}^{l-1}$, $\hat{\mathbf{C}}_{\mathrm{out}}^{l}$ and $\mathbf{C}_{\mathrm{out}}^{l}$ represent the depthwise convolution of the l^{th} layer, HSTM and MLP outputs, respectively.

Based on these, our model focus only on sequences that are highly relevant to the current query in a query-adaptive way, while avoiding distraction caused by irrelevant sequences, thus reducing the computation and minimizing the feature redundancy without sacrificing performance.

C. Cross-Temporal Collaborative Attention

Before calculating the true differences between dual-temporal RS images, it is necessary to deeply explore the interactions between features to avoid irrelevant disturbances such as seasonal changes, variations in lighting angles, and building renovations. We design a CTCA module which facilitates attention allocation to truly changed regions through interactive learning between dual-temporal RS images.

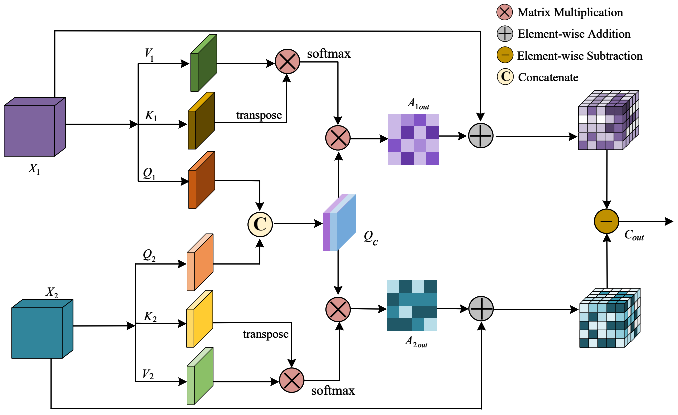

Fig. 3. The structure of CTCA module.

To strike a balance between memory usage and computational efficiency, we introduce the idea of efficient attention (EA) [10]. The computational complexity is reduced to a linear level by simply changing the matrix multiplication order of the original SA mechanism, achieving more efficient computation.

$$\mathbf{A}_{\text{eff}} = \text{Softmax}(\frac{\mathbf{K}^{\top}\mathbf{V}}{\sqrt{d_k}})\mathbf{Q} \tag{5}$$

where the scalar factor $\sqrt{d_k}$ is introduced to avoid concentrated weights and gradient vanishing.

Based on the above background, we design a CTCA module, as shown in Fig. 3. The CTCA module comprehensively considers the special characteristics of time span and target changes. Channel fusion and EA are performed on the $\mathbf{Q}$ of dual-temporal RS images.

$$\mathbf{C}_{\text{out}} = \text{Sub}((\mathbf{A}_{1\text{out}} + \mathbf{X}_1), (\mathbf{A}_{2\text{out}} + \mathbf{X}_2)) \tag{6}$$

where $\mathbf{A}_{1out}$ and $\mathbf{A}_{2out}$ are calculated by (5).

The CTCA module can support single-temporal RS images to fuse feature representations of another temporal while preserving their own features, overcoming semantic differences between dual-temporal RS images. Through a clever channel fusion mechanism, true differential features can be better captured from single-temporal RS images. In this way, it effectively explores the feasibility of interaction between dual-temporal images, which not only improves the sensitivity of the model to changed regions, but also reduces the influence of irrelevant factors.

III. Experiments

To further verify the effectiveness of the proposed method in high-resolution RS image CD, experiments are conducted on two public CD datasets: the LEVIR-CD dataset [11] and the DSIN-CD dataset [12].

A. Training Details

The experiments are conducted using the deep learning framework PyTorch. We implemented the proposed method with the batch size set to 32 for 200 epochs on the device with an NVIDIA GeForce RTX 3090 GPU, and the initial learning rate is 0.0001. The Adam optimizer is used to optimize the model with momentum of 0.99 and weight decay of 0.0005. We use a combination of BCE Loss and Dice Loss to optimize network weights.

B. Evaluation and Results

To evaluate the superiority of the proposed method, the DSECTCA-Net is compared with 11 state-of-the-art CD methods, which can be roughly categorized into two groups. Firstly, CNN-based methods: FCN-PP [13], STA Net [11], FDCNN [14], SNUNet [15], IF-Net [12], and DSAMNet [16]. Secondly, Transformer-based methods: BIT [3], Hybrid-TransCD [17], SwinSUNet [18], ChangeFormer [19], and WNet [20].

Quantitative Evaluation: To verify the effectiveness of the proposed method, we conducted the comparative experiments, mainly using three metrics for comprehensive evaluation of the proposed method, including Precision (Pre), Recall (Rec), and the harmonic index F1-score. As shown in Table I, the best values of the experimental results are shown in bold and the second-best are underlined. On the LEVIR-CD dataset, the CD performance of our proposed method surpasses the second-ranked method, ChangeFormer, by 0.46% with 41.11% of the number of its parameters and 4.69% of its computation. While on the DSIFN-CD dataset, the CD performance of the proposed method outperforms the second-ranked method, SwinSUNet, by 1.04% with 33.09% of its parameters and 44.88% of its computation. The CD performance of methods with smaller number of parameters or computation amount, such as STA-Net, BIT, etc., generally performed less effectively on the both datasets. These results fully prove the effectiveness of our proposed method, which is able to effectively reduce the number of parameters and computation amount of the model while ensuring the feature extraction capability, and achieve better detection results.

Qualitative Evaluation: To show the significant advantages of the proposed method, we selected representative samples for visual comparison, as shown in Fig. 4. In the yellow box in the first row of the figure, RS images may exhibit similar behavior between target objects and backgrounds due to changes in lighting angles. Methods such as STA-Net and FDCNN may mistakenly learn shadow areas resulting in serious false detection. However, when introducing time factors, DSECTCA-Net can alleviate the problem of shadow interference. It not only effectively suppresses irrelevant factors, but also exhibits excellent internal integrity of changed objects.

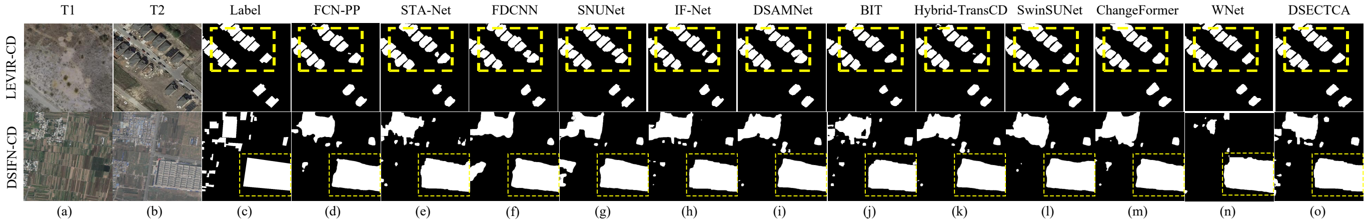

Fig. 4. Comparative experiments on two public datasets. (a) T1 images. (b) T2 images. (c) labels. (d) FCN-PP. (e) STA-Net. (f) FDCNN. (g) SNUNet. (h) IF-Net. (i) DSAMNet. (j) BIT. (k) Hybrid-TransCD. (l) SwinSUNet. (m) ChangeFormer. (n) WNet. (o) Ours.

TABLE I
QUANTITATIVE COMPARISONS OF DIFFERENT METHODS

Method Type	Number	Network	LEVIR-CD			DSIFN-CD			Params(M)	FLOPs(G)
			Pre(%)	*Rec*(%)	*F1*(%)	*Pre*(%)	*Rec*(%)	*F1*(%)		
CNN	1	FCN-PP [13]	80.31	89.48	84.64	56.42	59.25	57.80	28.13	34.65
	2	STA-Net [11]	86.17	89.39	87.73	66.22	67.16	66.69	16.93	**6.58**
	3	FDCNN [14]	82.99	88.71	85.76	64.42	68.38	66.34	13.71	32.40
	4	SNUNet [15]	89.06	87.53	88.29	62.47	69.74	65.90	12.03	33.04
	5	IF-Net [12]	89.73	86.06	87.80	72.36	63.86	67.85	50.71	41.18
	6	DSAMNet [16]	82.75	88.39	85.48	61.28	75.41	67.62	16.95	75.29
Transformer	7	BIT [3]	89.24	89.37	89.31	68.36	70.18	69.26	**6.93**	8.44
	8	Hybrid-TransCD [17]	91.45	88.72	90.06	68.79	70.42	69.69	166.57	51.38
	9	SwinSUNet [18]	90.51	89.72	90.11	68.72	71.68	70.17	50.95	21.19
	10	ChangeFormer [19]	92.05	88.80	90.40	69.38	70.51	69.94	41.01	202.83
	11	WNet [20]	90.48	90.17	90.33	68.85	69.03	68.94	42.56	19.20
ours	12	DSECTCA-Net	90.52	91.21	**90.86**	70.96	71.46	**71.21**	16.86	9.51

C. Ablation Experiments

The foundational network framework of the proposed method uses the original four-stage Transformer network as the encoding part, and the decoding part uses asymmetric ghost convolution to replace the original 3×3 convolution, and the difference maps obtained by subtracting feature maps from each layer are connected by skip connections. Table II shows a series of ablation experiments of DSECTCA-Net on the LEVIR-CD dataset.

TABLE II
ABLATION EXPERIMENTS

Methods	*Pre*(%)	*Rec*(%)	*F1*(%)	Params(M)	FLOPs(G)
Base	89.20	89.52	89.36	25.92	12.51
Base+CTCA	89.95	89.67	89.81	25.94	13.35
Base+DSE	89.95	90.64	90.29	16.28	9.15
Base+DSE+CTCA	90.52	91.21	**90.86**	16.86	9.51

After combining the DSE module and the CTCA module, the detection accuracy F1 is improved by 1.50%, while the amount of parameters is reduced by 9.06M, and the amount of computation is reduced by 3.0G compared to the base network. This fully proves the effectiveness of the proposed modules.

By replacing the encoder in the base network with the dynamic sparse encoder, and reducing the feature correlation computation through the HSTM in the DSE module, the F1 is improved by 0.93% over the baseline. Additionally, the parameter of the model as well as the computational complexity is effectively reduced compared to the base network.

By adding the CTCA module to the base network, our network fully models the dual-temporal RS images by introducing temporal information, which enables the real difference features captured from single-temporal RS images, effectively explores the feasibility of the interaction between the dual-temporal. This not only improves the sensitivity of the model to the changing regions, but also reduces the interference of irrelevant factors to a certain extent.

IV. CONCLUSION

In this paper, we have proposed a DSECTCA-Net method for CD tasks. First, a DSE module is designed to reduce the computation of feature correlation and irrelevant sequences by using HSTM, being able to extract features dynamically and adaptively. Secondly, a CTCA module is introduced to model the temporal concepts between dual-temporal RS images by fully exploring the feasibility of interaction, and shifting the attention to the real changed features so as to reduce the occurrence of false detection and missed detection. Finally, the analysis of full comparative experiments and ablation experiments on the two public CD datasets is conducted to further demonstrates the effectiveness of our DSECTCA-Net.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (Program No. 62271296, 62201334), in part by Scientific Research Program Funded by Shaanxi Provincial Education Department (23JP014, 23JP022).

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/ICASSP49660.2025.10890539, ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)Demo

References

[1] Chen Wu, Bo Du, and Liangpei Zhang, ”Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9774-9788, 2023.

[2] Tao Lei, Jie Wang, Hailong Ning, Xingwu Wang, Dinghua Xue, Qi Wang, and Asoke K. Nandi, ”Difference enhancement and spatial-spectral nonlocal network for change detection in VHR remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-13, 2022.

[3] Hao Chen, Zipeng Qi, and Zhenwei Shi, ”Remote sensing image change detection with transformers,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2022.

[4] Tao Lei, Xinzhe Geng, Hailong Ning, Zhiyong Lv, Maoguo Gong, Yaochu Jin, and Asoke K. Nandi., ”Ultralightweight spatial-spectral feature cooperation network for change detection in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-14, 2023.

[5] Dinghua Xue, Tao Lei, Shuangming Yang, Zhiyong Lv, Tongfei Liu, Yaochu Jin, and Asoke K. Nandi, ”Triple change detection network via joint multifrequency and full-scale swin-transformer for remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15, 2023.

[6] Tao Lei, Yetong Xu, Hailong Ning, Zhiyong Lv, Chongdan Min, Yaochu Jin, and Asoke K. Nandi, ”Lightweight structure-aware transformer network for remote sensing image change detection,” *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1-5, 2024.

[7] Zhiyong Lv, Jie Liu, Weiwei Sun, Tao Lei, Jón Atli Benediktsson, and Xiuping Jia, ”Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15, 2023.

[8] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, ”Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 548-558.

[9] Kun Dai, Ke Wang, Tao Xie, Tao Sun, Jinhang Zhang, Qingjia Kong, Zhiqiang Jiang, Ruifeng Li, Lijun Zhao, and Mohamed Omar, ”DSAP: Dynamic sparse attention perception matcher for accurate local feature matching,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1-16, 2024.

[10] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li, ”Efficient attention: Attention with linear complexities,” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3530-3538.

[11] Hao Chen and Zhenwei Shi, ”A spatial-temporal attention-based method and a new dataset for remote sensing image change detection”, *Remote Sensing*, vol. 12, no. 10, pp. 1662, 2020.

[12] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang, ”A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16, 2022.

[13] Tao Lei, Yuxiao Zhang, Zhiyong Lv, Shuying Li, Shigang Liu, and Asoke K. Nandi, ”Landslide inventory mapping from bitemporal images using deep convolutional neural networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 982-986, 2019.

[14] Min Zhang and Wenzhong Shi, ”A feature difference convolutional neural network-based change detection method,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7232-7246, 2020.

[15] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li, ”SNUNet-CD: A densely connected siamese network for change detection of VHR images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022.

[16] Qingtian Ke and Peng Zhang, ”Hybrid-TransCD: A hybrid transformer remote sensing image change detection network via token aggregation”, *ISPRS International Journal of Geo-Information*, vol. 11, no. 4, pp. 263, 2022.

[17] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu, ”A deeply supervised image fusion network for change detection in high-resolution bi-temporal remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183-200, 2020.

[18] Cui Zhang, Liejun Wang, Shuli Cheng, and Yongming Li, ”SwinSUNet: Pure transformer network for remote sensing image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-13, 2022.

[19] Wele Gedara Chaminda Bandara and Vishal M. Patel, ”A transformer-based siamese network for change detection,” *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 207-210.

[20] Xu Tang, Tianxiang Zhang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao, ”WNet: W-Shaped hierarchical network for remote-sensing image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-14, 2023.