Global and Local Attention-Based Transformer for Hyperspectral Image Change Detection

Ziyi Wang, Feng Gao, *Member, IEEE*, Junyu Dong, *Member, IEEE*, and Qian Du, *Fellow, IEEE*

***Abstract*— Recently, transformer-based hyperspectral image (HSI) change detection methods have shown remarkable performance. Nevertheless, existing attention mechanisms in transformers have limitations in local feature representation. To address this issue, we propose global and local attention-based transformer (GLAFormer), which incorporates a global and local attention module (GLAM) to combine high-frequency and low-frequency signals. Furthermore, we introduce a cross-gating mechanism, called cross-gated feedforward network (CGFN), to emphasize salient features and suppress noise interference. Specifically, the GLAM splits attention heads into global and local attention components to capture comprehensive spatial–spectral features. The global attention component uses global attention on downsampled feature maps to capture low-frequency information, while the local attention component focuses on high-frequency details using nonoverlapping window-based local attention. The CGFN enhances the feature representation via convolutions and cross-gating mechanism in parallel paths. The proposed GLAFormer is evaluated on three HSI datasets. The results demonstrate its superiority over state-of-the-art HSI change detection methods. The source code of GLAFormer is available at https://github.com/summitgao/GLAFormer.**

***Index Terms*— Change detection, gating mechanism, global and local attention, hyperspectral image (HSI), vision transformer.**

I. Introduction

HYPERSPECTRAL image (HSI) change detection stands as a crucial task within the field of remote sensing, focusing on the identification of altered areas by comparing hyperspectral images obtained at different times. The exceptional spectral resolution of HSIs facilitates accurate detection of changes in ground objects [1]. As such, HSI change detection has been widely applied in various domains, including damage assessment [2], land cover analysis [3], and urban expansion monitoring [4].

Traditional methods in HSI change detection primarily used techniques, such as change vector analysis [5] and Tucker decomposition [6], to analyze the spectral changes in multitemporal images. Niemeyer and Canty [7] were the first to introduce the multivariate alteration detection (MAD) method, which is based on canonical correlation analysis and is designed for unsupervised change detection in vegetation using multitemporal hyperspectral images. Later, Nielsen [8] enhanced this approach by developing the iteratively reweighted MAD (IR-MAD) algorithm. However, these methods encountered limitations in threshold selection and reduced robustness in complex scenarios [9]. Recently, convolutional neural networks (CNNs) [10] have been proven especially effective in extracting representative features from multitemporal HSIs. Saha et al. [11] proposed a method called deep CVA by combining CNNs with change vector analysis (CVA).

More recently, transformers, which rely entirely on self-attention, have gained popularity in computer vision tasks, such as image classification [12] and object detection [13]. Transformers have the advantage of capturing global dependencies and exhibit better performance in handling long-range dependencies compared with CNNs. This potential has prompted researchers to explore the attention mechanisms for HSI change detection. Song et al. [14] enhanced the feature representation of multitemporal HSIs by introducing cross-temporal interaction symmetric attention. Furthermore, Ding et al. [1] introduced the transformer encoder for HSI change detection, leveraging self-attention with a global receptive field to enhance the recognition of changes. In [15], a transformer-based multiscale feature fusion model was proposed for change detection.

Although the existing transformer-based methods for HSI change detection have achieved promising performance, they still suffer from two limitations.

1) *Insufficient local-level representations modeling:* Transformers tend to pay more attention to global features. However, global and local features serve distinct roles in encoding HSI patterns. The emphasis on global features in transformers leads to the loss of certain local features, which results in a degradation of change detection performance.
2) *Limited nonlinear feature transformation:* Feedforward network (FFN) is commonly used to process the output from the attention layer in transformer, enabling nonlinear feature transformation for the input of the subsequent attention layer. However, the existing methods are limited in nonlinear feature representation and are susceptible to noise interference.

To overcome the above limitations, we propose a global and local attention-based transformer (GLAFormer) for

Received 21 August 2024; accepted 20 November 2024. Date of publication 25 November 2024; date of current version 11 December 2024. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0117201 and in part by the Natural Science Foundation of Qingdao under Grant 23-2-1-222-ZYYD-JCH. *(Corresponding author: Feng Gao.)*

Ziyi Wang, Feng Gao, and Junyu Dong are with the School of Information Science and Engineering, Ocean University of China, Qingdao 266100, China (e-mail: gaofeng@ouc.edu.cn).

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 USA.

Digital Object Identifier 10.1109/LGRS.2024.3505294

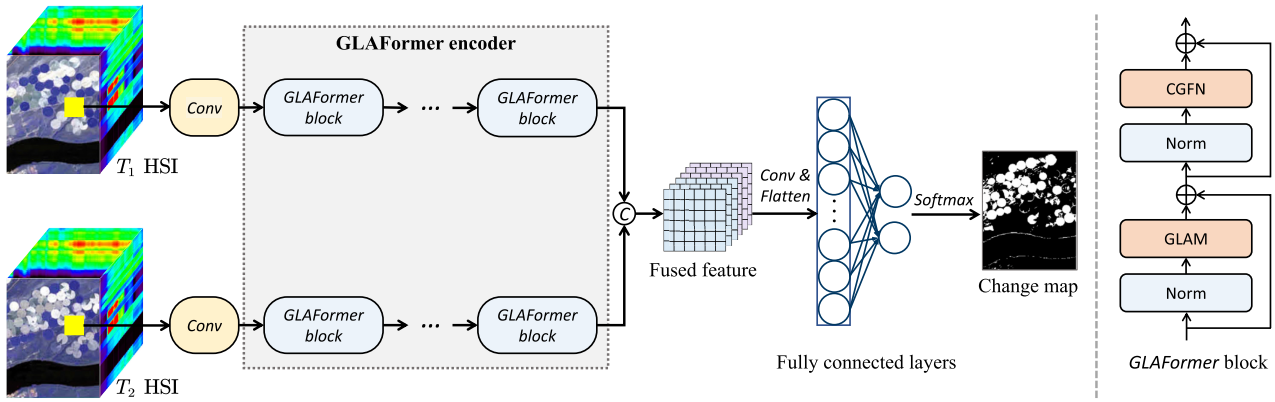

Fig. 1. Overview of the proposed GLAFormer. In the GLAFormer block, high-frequency and low-frequency signals are fused through GLAM. The ability to capture local context is enhanced through the CGFN. The final output is the change map, where white pixels represent the changed areas.

HSI change detection, GLAFormer for short. Specifically, to enhance the local-level feature representations, we have designed a global and local attention module (GLAM) to encode both low-frequency and high-frequency signals. Furthermore, to augment the nonlinear feature transformation, we propose the cross-gated feedforward network (CGFN) to amplify the salient information while suppressing noise. Extensive experiments on three HSI change detection datasets demonstrate the superiority of our proposed GLAFormer.

Our main contributions can be summarized as follows.

1) We propose the GLAM as an enhancement to the self-attention mechanism. This module combines high-frequency and low-frequency signals to achieve a more comprehensive spatial–spectral feature representation for change detection.
2) We develop the CGFN to improve the nonlinear feature transformation within transformers. This network amplifies important information and mitigates noise interference.
3) Extensive experimental results demonstrate that the proposed GLAFormer outperforms state-of-the-art methods. The codes will be released to the remote sensing community.

II. Methodology

The overall architecture of our proposed GLAFormer is shown in Fig. 1. Two hyperspectral images (T_1 and T_2) captured at different times are passed to the GLAFormer. First, two patches from the multitemporal HSIs of the same geographical area are extracted. Then, both the patches are fed into two parallel GLAFormer encoders to extract informative and robust features. Next, the learned features from the two paths are fused. Finally, the fused features are transformed by several convolutional and fully connected layers for change detection.

As shown in the right part of Fig. 1, the GLAFormer block consists of two key modules: GLAM and CGFN, which are detailed as below.

A. Global and Local Attention Module

As depicted in Fig. 2, the GLAM consists of two branches: global attention and local attention. The global attention captures the global dependencies of the input, while the local attention branch computes the detailed local feature dependency. The global and local features are fused by concatenation.

1) Local Attention: The local attention branch encodes high-frequency features via local window self-attention, which applies self-attention mechanism to local windows of feature maps. As shown in Fig. 2, a local window refers to a 3×3 region in the feature maps. These local windows are evenly partitioned in a nonoverlapping manner.

The feature within each window is of size $s \times s \times d$. Here, s is the size of the window, and d is the feature dimension. The feature within each window is reshaped into $\mathbf{X}_\mathrm{l} \in \mathbb{R}^{d \times s^2}$. Next, a 1×1 convolution is applied to enhance the input and obtain the query $\mathbf{Q}_\mathrm{l}$, key $\mathbf{K}_\mathrm{l}$, and value $\mathbf{V}_\mathrm{l}$. The local window attention is defined as

$$\hat{\mathbf{X}}_\mathrm{l} = \mathbf{V}_\mathrm{l} \cdot \mathrm{Softmax}\left(\frac{\mathbf{K}_\mathrm{l}^T \mathbf{Q}_\mathrm{l}}{\sqrt{D}}\right) \tag{1}$$

where $\hat{\mathbf{X}}_\mathrm{l}$ is the output feature from the local attention branch. $\mathbf{Q}_\mathrm{l}$, $\mathbf{K}_\mathrm{l}$, and $\mathbf{V}_\mathrm{l}$ are the tensors after 1×1 convolutions specific to the local attention branch, and D is the number of hidden dimensions for a single head in the local attention.

2) Global Attention: The global attention branch captures low-frequency features by applying the attention mechanism over pooled feature maps. As illustrated in Fig. 2, the input feature is evenly partitioned into 3×3 windows in a nonoverlapping manner. To effectively capture global information, average pooling is used on each window to obtain the average-pooled feature map $\mathbf{Z} \in \mathbb{R}^{d \times 9}$. Next, $\mathbf{Z}$ is transformed to key $\mathbf{K}_\mathrm{g} \in \mathbb{R}^{d \times 9}$ and value $\mathbf{V}_\mathrm{g} \in \mathbb{R}^{d \times 9}$. To ensure complete and unchanged information access, the global attention uses queries $\mathbf{Q}_\mathrm{l} \in \mathbb{R}^{d \times 81}$ from the original feature map. This approach is consistent with that of local attention. To generate the output features $\hat{\mathbf{Z}}_\mathrm{g}$, the standard self-attention is applied on $\mathbf{Q}_\mathrm{l}$, $\mathbf{K}_\mathrm{g}$, and $\mathbf{V}_\mathrm{g}$

$$\hat{\mathbf{Z}}_\mathrm{g} = \mathbf{V}_\mathrm{g} \cdot \mathrm{Softmax}\left(\frac{\mathbf{K}_\mathrm{g}^T \mathbf{Q}_\mathrm{l}}{\sqrt{D}}\right). \tag{2}$$

3) Channel Splitting and Merging: The input features are evenly split along the channel dimension before entering the global and local attention branches, reducing complexity and

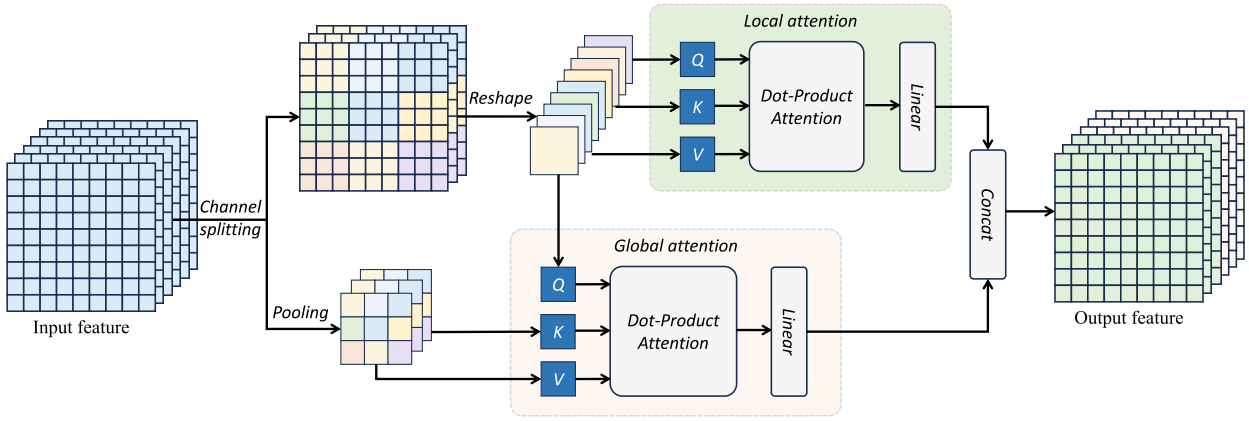

Fig. 2. Illustration of GLAM. The module consists of two branches: global attention and local attention. The global attention branch captures the long-range dependencies of the input, while the local attention branch computes the detailed local feature dependency. Then, the global and local features are fused by concatenation.

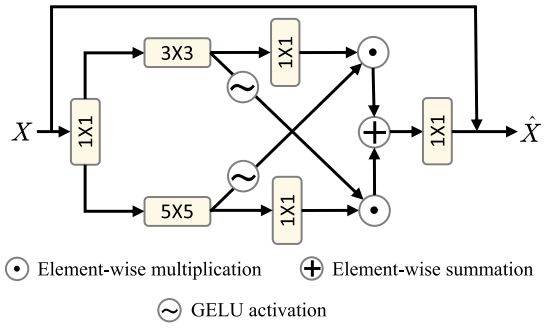

Fig. 3. Architecture of CGFN.

boosting GPU throughput. This splitting also decomposes the learnable parameters into smaller matrices, reducing the model's parameter count. These two sets of features are separately fed into the local and global attention branches, respectively. To produce the output of GLAM, the output features $\hat{\mathbf{X}}_{\mathrm{l}}$ from the local attention branch and the output features $\hat{\mathbf{Z}}_{\mathrm{g}}$ from the global attention branch are concatenated as

$$\mathbf{F}_{\mathrm{GLAM}} = \mathrm{concat}\left[\hat{\mathbf{X}}_{\mathrm{l}}; \hat{\mathbf{Z}}_{\mathrm{g}}\right]. \tag{3}$$

B. Cross-Gated Feedforward Network

To enhance the nonlinear feature transformation in transformers, the CGFN is proposed, which incorporates the gating mechanism and multiscale convolution into the existing feedforward network.

As shown in Fig. 3, the proposed CGFN consists of two parallel paths. In each path, depthwise convolutions with different sizes of kernels are used to enhance the multiscale feature extraction. Then, the gating mechanism is used to filter the less informative features in each path. The useful features passing through the gates are fused with the original features from another path. The fused features from the two paths are combined via elementwise summation. Given input $\mathbf{Y}$, the CGFN can be defined as

$$\begin{aligned}\mathrm{Gating}(\mathbf{Y}) = {} & \phi\left(W_{3\times3}W^{0}_{1\times1}\mathbf{Y}\right) \odot \left(W^{2}_{1\times1}W_{5\times5}W^{0}_{1\times1}\mathbf{Y}\right) \\ & + \phi\left(W_{5\times5}W^{0}_{1\times1}\mathbf{Y}\right) \odot \left(W^{3}_{1\times1}W_{3\times3}W^{0}_{1\times1}\mathbf{Y}\right) \quad (4)\end{aligned}$$

$$\hat{\mathbf{Y}} = W^{4}_{1\times1}\mathrm{Gating}(\mathbf{Y}) + \mathbf{Y} \tag{5}$$

where $\hat{\mathbf{Y}}$ denotes the output features, Gating$(\cdot)$ denotes the cross-gated mechanism, ϕ is the GELU activation function, and $\odot$ represents the elementwise multiplication operation. The gating mechanism and the multiscale convolutions amplify important information and mitigate noise interference.

III. Experimental Results and Analysis

A. Dataset and Experimental Setting

We evaluate the performance of our GLAFormer through extensive experiments on three widely recognized multitemporal hyperspectral datasets. These datasets were sourced from the Hyperion sensor onboard the EO-1 satellite. Specifically, the first dataset, referred to as the River dataset [16], comprises imagery of a river in Jiangsu Province, China. The second dataset is the Farmland dataset [17], which covers a farmland area in Yancheng City, Jiangsu Province, China. The third, known as the Hermiston dataset [9], captures irrigated farmland in Hermiston City, Umatilla County, Oregon, USA.

To demonstrate the effectiveness of the proposed GLAFormer, six state-of-the-art models are selected for comparison, i.e., IR-MAD [8], SSA-SimaNet [18], SSCNN-S [19], CDFormer [1], SSTFormer [20], CSDBF [21], and GTMS-iam [22]. GLAFormer is configured with four blocks and eight attention heads across all three datasets. The comparative analysis is grounded in two primary metrics: overall accuracy (OA) and the Kappa coefficient. OA provides a general assessment of change detection performance in terms of overall correctness. Kappa is a more robust measure that takes into account the agreement between the observed classification and what would be expected by chance.

All the experiments are carried out using the PyTorch framework. The training phase spanned over 100 epochs and was conducted on a single NVIDIA 4090 GPU. The Adam optimizer is used with a learning rate of 0.0006. The training batch size is set as 128. The input patch size for the proposed methods is 9×9, and the dimension of the embedded sequence is fixed at 256. For each of the three datasets, 3% of the samples are selected for training, 2% for validation, and the remaining samples are used for testing.

TABLE I

QUANTITATIVE EVALUATION OF DIFFERENT CHANGE DETECTION METHODS ON THREE DATASETS. THE BEST PERFORMER IS MARKED IN BLUE AND THE SECOND BEST IS UNDERLINED. THE HIGHER MEANS BETTER PERFORMANCE

Dataset	River		Farmland		Hermiston	
Measure	OA	Kappa	OA	Kappa	OA	Kappa
IR-MDA	94.07	62.96	95.97	90.13	86.75	57.86
SSA-SiamNet	94.87	72. 43	94.79	87. 43	93. 93	83.19
SSCNN-S	95. 53	75.67	95.49	88. 85	95.69	87.08
CDFormer	94.92	72.76	97.34	93.58	93.06	81.57
SSTFormer	96.72	81.63	98.01	95.14	93.39	82.73
CSDBF	95.56	75.61	98.23	95.75	96.13	89.51
GTMSiam	97.11	82.93	97.37	94.03	95.66	87.26
GLAFormer	**97.81**	**83.72**	**98.95**	**97.14**	**97.23**	**91.68**
	↑0.70%	↑0.78%	↑1.58%	↑3.11%	↑1.57%	↑4.42%

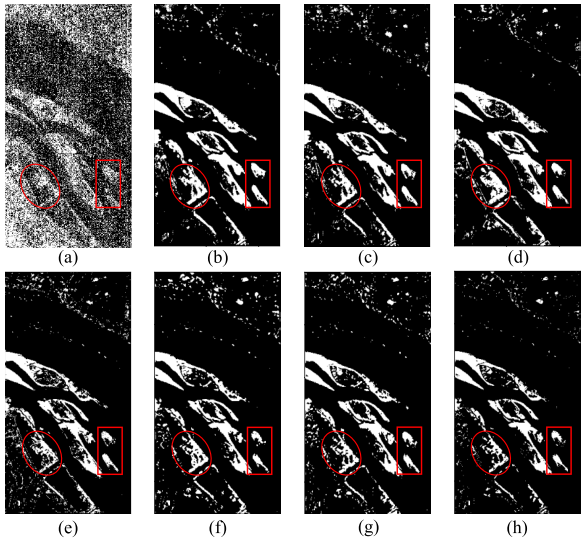

Fig. 4. Change detection results of different methods on the River dataset. (a) IR-MAD. (b) SSA-SiamNet. (c) SSCNN-S. (d) CDFomer. (e) SSTFormer. (f) CSDBF. (g) Proposed GLAFormer. (h) Ground truth.

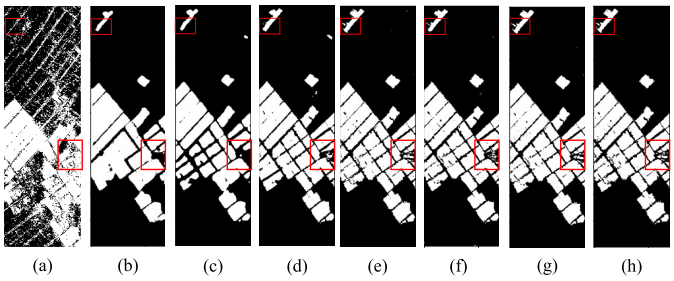

Fig. 5. Change detection results of different methods on the Farmland dataset. (a) IR-MAD. (b) SSA-SiamNet. (c) SSCNN-S. (d) CDFomer. (e) SSTFormer. (f) CSDBF. (g) Proposed GLAFormer. (h) Ground truth.

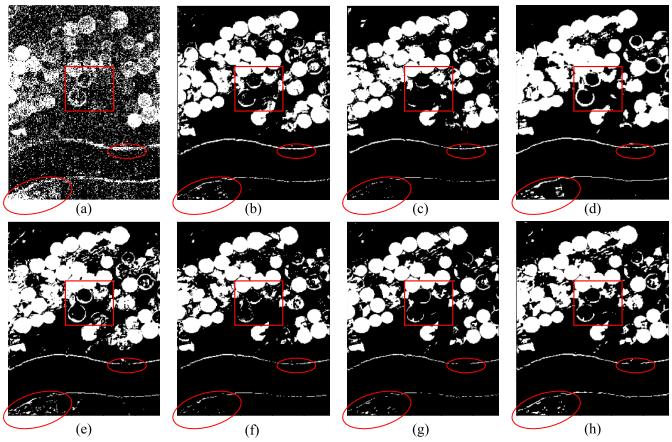

Fig. 6. Change detection results of different methods on the Hermiston dataset. (a) IR-MAD. (b) SSA-SiamNet. (c) SSCNN-S. (d) CDFomer. (e) SSTFormer. (f) CSDBF. (g) Proposed GLAFormer. (h) Ground truth.

B. Experimental Results and Comparison

As presented in Table I, the quantitative comparison between the GLAFormer and other methods is conducted on three datasets. The results demonstrate that our proposed method consistently outperforms the compared methods across all three datasets in terms of OA and Kappa. Notably, in terms of the Kappa coefficient, GLAFormer achieves an average improvement of 2.77% across the three datasets compared with the previous state-of-the-art feature fusion method, GTMSiam. This signifies an accuracy boost of over 20% in regions that were challenging for previous models to identify.

To demonstrate the superior performance of our proposed GLAFormer, we also qualitatively compare the results of different methods on the three datasets. Fig. 4 shows the change detection result on the River dataset. Unlike other methods, the IR-MAD approach exhibits significant noise due to its lack of training information in change detection. In contrast, the detection results of GLAFormer are closer to the ground truth, particularly in the highlighted regions.

The change detection results of the Farmland dataset are shown in Fig. 5. The regions in the red rectangle reveal that only the GLAFormer and SSTFormer successfully identify the subtle alterations within the region. This observation aligns with the quantitative findings reported in Table I. It is worth noting that GLAFormer demonstrates an improvement of 1.58% and 3.11% in OA and Kappa coefficient, respectively, compared with the GTMSiam method. This performance gain can be attributed to superior integration of global and local signals in GLAFormer.

The change detection results of the Hermiston dataset are depicted in Fig. 6. It shows the complex nature of the changes within this dataset, characterized by numerous irregular regions. A detailed analysis of the highlighted regions reveals that our proposed method produces a more accurate change map with clearer boundaries and reduced noise, underscoring its robust capability in detecting changes within complex backgrounds. This performance improvement can be attributed to the effectiveness of the dual-gated mechanism of CGFN in noise suppression. Quantitatively, GLAFormer achieves a 1.57% and 4.42% increase in OA and Kappa coefficient, respectively, compared with the GTMSiam method.

Both the qualitative and quantitative analyses demonstrate the robustness and accuracy of the proposed GLAFormer in handling complex change detection scenarios.

C. Ablation Study

We conduct a series of ablation experiments on the three datasets to validate the effectiveness of the proposed GLAM

TABLE II
ABLATION STUDIES OF THE PROPOSED GLAFORMER. THE BEST PERFORMER IS MARKED IN BLUE

Method	OA on different datasets (%)		
	River	Farmland	Hermiston
Basic Transformer	97.19	97.87	94.42
GLAFormer w/o GLAM	97.17	98.54	95.20
GLAFormer w/o CGFN	97.64	98.66	96.97
Proposed GLAFormer	97.81	98.95	97.23

and CGFN. First, we design a basic transformer with the same structure as the GLAFormer, while the basic transformer uses traditional multihead attention. In addition, we have two variants of GLAFormer, i.e., without the GLAM (w/o GLAM) and without the CGFN (w/o CGFN). These are replaced with standard self-attention and FFN, respectively. The results of the ablation study are shown in Table II. We find that GLAFormer and its variants beat the basic transformer in all cases. The GLAFormer always achieves better performance than its two variants on the three datasets. This demonstrates the necessity of the GLAM and CGFN designed in GLAFormer.

IV. CONCLUSION

In this letter, we propose a novel GLAFormer for HSI change detection. The GLAFormer offers two enhancements over existing transformer-based change detection methods. First, the designed GLAM leverages the abundant channel information intrinsic to hyperspectral images to combine both high-frequency and low-frequency signals. Furthermore, the designed CGFN is meticulously engineered to augment the extraction of pertinent information while concurrently mitigating noise interference, thereby enhancing the overall quality of the change detection process. Our comprehensive experiments conducted on three hyperspectral datasets consistently demonstrate the superior performance of GLAFormer over the state-of-the-art methods.

REFERENCES

[1] J. Ding, X. Li, and L. Zhao, "CDFormer: A hyperspectral image change detection method based on transformer encoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[2] P. K. E. Campbell et al., "Detection of initial damage in Norway spruce canopies using hyperspectral airborne data," *Int. J. Remote Sens.*, vol. 25, no. 24, pp. 5557–5584, Dec. 2004.

[3] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[4] J. Yuan, S. Wang, C. Wu, and Y. Xu, "Fine-grained classification of urban functional zones and landscape pattern analysis using hyperspectral satellite imagery: A case study of Wuhan," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3972–3991, 2022.

[5] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2007.

[6] Z. Hou, W. Li, R. Tao, and Q. Du, "Three-order Tucker decomposition and reconstruction detector for unsupervised hyperspectral change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6194–6205, 2021.

[7] I. Niemeyer and M. J. Canty, "Pixel-based and object-oriented change detection analysis using high-resolution imagery," in *Proc. 25th Symp. Safeguards Nucl. Mater. Manage.*, 2003, pp. 2133–2136.

[8] A. A. Nielsen, "The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[9] M. Hasanlou and S. T. Seydi, "Hyperspectral change detection: An experimental comparative study," *Int. J. Remote Sens.*, vol. 39, no. 20, pp. 7029–7083, Oct. 2018.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

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

[12] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–12.

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

[14] R. Song, W. Ni, W. Cheng, and X. Wang, "CSANet: Cross-temporal interaction symmetric attention network for hyperspectral image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] Z. Wang, H. Luo, P. Wang, F. Ding, F. Wang, and H. Li, "VTC-LFC: Vision transformer compression with low-frequency components," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 13974–13988.

[16] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2018.

[17] Y. Yuan, H. Lv, and X. Lu, "Semi-supervised change detection method for multi-temporal hyperspectral images," *Neurocomputing*, vol. 148, pp. 363–375, Jan. 2015.

[18] L. Wang, L. Wang, Q. Wang, and P. M. Atkinson, "SSA-SiamNet: Spectral–spatial-wise attention-based Siamese network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5510018.

[19] T. Zhan et al., "SSCNN-S: A spectral–spatial convolution neural network with Siamese architecture for change detection," *Remote Sens.*, vol. 13, no. 5, p. 895, Feb. 2021.

[20] Y. Wang et al., "Spectral–spatial–temporal transformers for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536814.

[21] X. Wang, K. Zhao, X. Zhao, and S. Li, "CSDBF: Dual-branch framework based on temporal–spatial joint graph attention with complement strategy for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540118.

[22] X. Wang, K. Zhao, X. Zhao, and S. Li, "GTMSiam: Gated transmitting-based multiscale Siamese network for hyperspectral image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.