Convolution and Attention Mixer for Synthetic Aperture Radar Image Change Detection

Haopeng Zhang, Zijing Lin, Feng Gao, Junyu Dong, Qian Du, and Heng-Chao Li

Abstract—**Synthetic aperture radar (SAR) image change detection is a critical task and has received increasing attentions in the remote sensing community. However, existing SAR change detection methods are mainly based on convolutional neural networks (CNNs), with limited consideration of global attention mechanism. In this letter, we explore Transformer-like architecture for SAR change detection to incorporate global attention. To this end, we propose a convolution and attention mixer (CAMixer). First, to compensate the inductive bias for Transformer, we combine self-attention with shift convolution in a parallel way. The parallel design effectively captures the global semantic information via the self-attention and performs local feature extraction through shift convolution simultaneously. Second, we adopt a gating mechanism in the feed-forward network to enhance the non-linear feature transformation. The gating mechanism is formulated as the element-wise multiplication of two parallel linear layers. Important features can be highlighted, leading to high-quality representations against speckle noise. Extensive experiments conducted on three SAR datasets verify the superior performance of the proposed CAMixer. The source codes will be publicly available at https://github.com/summitgao/CAMixer.**

Index Terms—**Change detection; Synthetic aperture radar; Shift convolution; Gating mechanism.**

I. INTRODUCTION

SYNTHETIC aperture radar (SAR) image change detection is widely acknowledged as a fundamental task in interpreting and understanding remote sensing data. It has significant implications for various applications, including land cover monitoring such as land cover monitoring [1] [2], and disaster monitoring [3] [4] [5]. With the increasing availability of multitemporal SAR images, the development of reliable change detection methods applicable to real-world scenarios has become crucial [6].

While many supervised and unsupervised methods have been proposed for SAR change detection, supervised methods often require prior knowledge and high-quality labeled samples, which are inconvenient or even difficult to collect in real applications. Furthermore, existing unsupervised methods are commonly based on convolutional neural networks (CNNs),

Haopeng Zhang, Zijing Lin, Feng Gao, and Junyu Dong are with the School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China. *(Corresponding author: Feng Gao)*

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

H. -C. Li is with the Sichuan Provincial Key Laboratory of Information Coding and Transmission, Southwest Jiaotong University, Chengdu 610031, China.

and have limitations in long-range feature modeling. Therefore, in this letter, we primarily focus on developing robust unsupervised SAR change detection method.

Recently, Liu et al. [7] introduced a spatial constraint on CNN. This spatial constraint restricts the convolution operations to local regions, thereby improving change detection performance. Saha et al. [8] proposed a Siamese convolutional network. This network employs a shared set of weights to handle multi-temporal SAR images. Wang et al. [9] employed a dual-path denoising network for SAR change detection. The network refines noise labels in training samples. Hafner et al. [10] employed a dual-stream U-Net and performed data fusion of Sentinel-1 and Sentinel-2 images. The fusion of multi-source data, along with the dual-stream architecture, enables accurate urban change detection. Liu et al. [11] proposed a change detection approach based on image translation. By transforming images of different types, it effectively detects changes from multi-source data, providing a versatile solution for unsupervised change detection.

Due to the inherent inductive bias in CNNs, existing methods possess the capability to discern subtle changes, such as edges and corners. Hence, the aforementioned CNN-based methods have demonstrated remarkable performance. However, with the emergence of Vision Transformer (ViT) [12], Transformer-based models have achieved significant success in various computer vision and image understanding tasks. These models utilize a global attention mechanism to capture long-range dependencies and compute informative features. Swin Transformer [13] achieves excellent performance in many vision tasks via shifted window self-attention computation. Despite their success, Transformers are rarely applied to multi-temporal SAR image analysis. Therefore, in this letter, we aim to investigate the potential of attention mechanism for SAR change detection task.

It is commonly non-trivial to design a robust Transformer-like framework for SAR change detection, since it possess the following challenges: 1) Transformers lack the inherent inductive bias of CNNs, making them less effective when training data is limited. 2) The non-linear transformation of the feed-forward network (FFN) has limitations in robust feature representation and is vulnerable to speckle noise.

To address these challenges, we present a **C**onvolution and **A**ttention **Mixer** for SAR change detection, **CAMixer** for short. First, to compensate the inductive bias for Transformer, we combine self-attention with shift convolution in a parallel way. The parallel design enriches feature representations by modeling convolution and attention simultaneously. Additionally, we adopt a gating mechanism in FFN to enhance the

Fig. 1. Illustration of the proposed Convolution and Attention Mixer (CAMixer) for SAR image change detection. The overall pipeline of CAMixer consists of three 3x3 convolutions and three mixing blocks. Each mixing block comprises a parallel Convolution and Attention Module (PCAM) and a Gated Feed-Forward Network (GFFN).

non-linear feature representations. The gating mechanism is formulated as element-wise multiplication of two parallel linear layers. Important features can be highlighted, leading to high-quality representations against the speckle noise.

In a nutshell, we summarize our contributions in threefold:

- We present a convolution and self-attention mixed network for SAR change detection. To the best of our knowledge, we are the first to explore the Transformer-like network for multi-temporal SAR data interpretation.
- We propose a gated feed-forward network (GFFN) for non-linear feature transformation. Gating mechanism is formulated as the element-wise product of two parallel paths of linear transformation layers, one of which is activated with the GELU activation. Hence, the GFFN selectively emphasizes important features, thereby mitigating the interference caused by speckle noise.
- Extensive experiments conducted on three SAR datasets demonstrate the effectiveness of the proposed CAMixer. In order to benefit other researchers, we have made our code publicly available.

II. METHODOLOGY

A. Framework of the Proposed CAMixer

SAR change detection aims to identify the changes that occur in the same area at different times (t_1 and t_2). The overview of CAMixer is shown in Fig. 1.

Preclassification is performed to generate training samples for CAMixer. Specifically, we first compute the difference image by the log-ratio operator. Then, hierarchical fuzzy c-means [14] are used to classify the difference image into changed, unchanged, and intermediate class. The pixels from changed and unchanged class are selected as training samples.

In the proposed CAMixer, several mixing blocks are employed for local and global feature extraction. Finally, the

extracted features are reshaped for classification. We now describe the key components of the mixing block: 1) Parallel Convolution and Attention Module (PCAM) and 2) Gated Feed-Forward Network (GFFN).

Fig. 2. Details of the PCAM. It consists of shift convolution and self-attention. The output of shift convolution and self-attention are fused through element-wise summation.

B. Parallel Convolution and Attention Module (PCAM)

As shown in Fig. 1, our PCAM is composed of shift convolution and self-attention.

Shift convolution. Inspired by Wang's work [15], we incorporate shift convolution for local feature extraction. It consists of a series of shift operations and a 1×1 convolution. The input features are evenly divided into five groups. The first four groups are shifted in different directions (left, right, top, bottom), while the last group remains unchanged.

In our implementation, we initially expand the number of channels of the input data X to βC using a 1×1 convolution, where β is the expansion ratio and C is the number of

Fig. 3. Relationship between the number of the parallel blocks and the PCC value.

Fig. 4. Visualization of the feature representations on the Chao Lake I dataset. (a) Features before the PCAM. (b) Features after the PCAM.

channels. Following the shift operation, we reduce the feature dimension back to the original size through another 1×1 convolution. This ensures consistency between the input and output feature sizes. Consequently, the shift convolution can be formulated as:

$$\hat{X} = W_{1\times1}^2(\text{shift}(W_{1\times1}^1(X))), \qquad (1)$$

where $W_{1\times1}^1$ is the first 1×1 convolution, and $W_{1\times1}^2$ is the second 1×1 convolution. Through the shift operation, channels of the input data are shifted, enabling cross-channel information fusion through channel mixing. The second 1×1 convolution leverages information from neighboring pixels, while the shift convolution facilitates the incorporation of large receptive fields, while maintaining a low computational burden.

Self-Attention Computation. Inspired by ViT [12], we first divided the image into non-overlapping patches (3×3 pixels), and encode each patch into a token embedding. Next, we compute query (Q), key (K), and value (V) via linear transformation of the token embedding. The output of self-attention is calculated by:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d})V, \qquad (2)$$

where $\sqrt{d}$ is a scaling factor. Finally, the output of shift convolution and self-attention are fused via element-wise summation. The obtained features are then normalized and fed into the GFFN to generate the input of the next mixing block.

C. Gated Feed-Forward Network

To enhance non-linear feature transformation, FFN is commonly used to process the output from the attention layer, enabling a better fit for the input of the subsequent attention layer. As illustrated in Fig. 1, we introduce the GFFN to further enhance representation learning. We make two modifications to the FFN: 1) multi-scale convolution and 2) gating mechanism. Firstly, we employ 3×3 and 5×5 depthwise convolutions to enhance the extraction of multi-scale information. Additionally, we utilize the gating mechanism

to emphasize the important components of the multi-scale convolutions.

The proposed GFFN is formulated as:

$$\hat{X} = W_{1\times1}^0\text{Gating}(X) + X, \qquad (3)$$

$$\text{Gating}(X) = \sigma(W_{1\times1}^1(X)) \odot \phi(X), \qquad (4)$$

$$\phi(X) = W_{3\times3}(W_{1\times1}^2(X)) + W_{5\times5}(W_{1\times1}^2(X)), \qquad (5)$$

where $W_{1\times1}^0, W_{1\times1}^1$, and $W_{1\times1}^2$ are 1×1 convolution. $W_{3\times3}$ denotes 3×3 depth-wise convolution, and $W_{5\times5}$ denotes 5×5 depth-wise convolution. Here, the $\odot$ is element-wise multiplication, and σ is the GeLU activation. To improve computational efficiency, we reduce the expansion ratio to 2 with marginal performance loss.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Datasets and Evaluation Metrics

We conducted experiments on three datasets, namely the Yellow River, Chao Lake I, and Chao Lake II datasets, to validate the effectiveness of the proposed CAMixer. The Yellow River dataset covers the Yellow River Estuary region in China, with images captured in June 2008 and June 2009 using the Radarset-2 SAR sensor. The Chao Lake I and II datasets cover a region of Chao Lake in China, with images captured in May 2020 and July 2020, respectively, using the Sentinel-1 sensor. During this period, Chao Lake experienced its highest recorded water level. The ground truth change maps for all three datasets were meticulously annotated by experts with prior knowledge.

To evaluate the performance of change detection, we employ five evaluation metrics: false positives (FP), false negatives (FN), overall error (OE), percentage of correct classification (PCC), and Kappa coefficient (KC).

B. Analysis of the Parallel Block Number

There are N PCAMs in the proposed CAMixer, and it is a critical parameter that may affect the change detection performance. To investigate the relationship between N and change detection accuracy, we set N from 0 to 8. Fig. 3 shows that when the number of PCAM increases, the value of PCC first increases and then becomes stable. However, more PCAM would increse the computational burden. Therefore, we set $N = 3$ for the Chao Lake II dataset, and $N = 5$ for the Yellow River and Chao Lake I datasets.

Fig. 5. Visualized results of different change detection methods on the three dataset: (a) Image captured at t_1. (b) Image captured at t_2. (c) Ground truth image. (d) Result by PCA-KM [16]. (e) Result by NR-CR [17]. (f) Result by NR-ELM [18]. (g) Result by DDNet [19]. (h) Result by MSAPNet [20]. (i) Result by the proposed CAMixer.

TABLE I
ABLATION STUDIES OF THE PROPOSED CAMIXER

Method	PCC on different datasets (%)		
	Yellow River	Chao Lake1	Chao Lake2
Basic Network	95.28	96.40	97.10
w/o PCAM	96.02	97.25	97.98
w/o GFFN	96.14	97.58	98.13
w/o H-Clustering	96.18	97.76	98.21
Ours	96.28	98.39	98.35

TABLE II
CHANGE DETECTION RESULTS OF DIFFERENT METHODS ON THREE
DATASETS

Method	Results on the Yellow River dataset				
	FP	FN	OE	PCC (%)	KC (%)
PCA-KM [16]	1835	2798	4633	93.76	78.34
NR-CR [17]	2257	2344	4601	93.80	79.03
NR-ELM [18]	629	3806	4435	94.03	77.80
DDNet [19]	1239	2161	3400	95.42	84.12
MSAPNet [20]	1206	2026	3232	95.65	84.96
Proposed CAMixer	619	2145	2764	96.28	86.86

Method	Results on the Chao Lake I dataset				
	FP	FN	OE	PCC (%)	KC (%)
PCA-KM [16]	12126	1786	13912	90.57	52.08
NR-CR [17]	2906	2892	5798	96.07	71.34
NR-ELM [18]	2282	3370	5652	96.17	70.71
DDNet [19]	3858	1182	5040	96.58	77.60
MSAPNet [20]	6632	1108	7740	94.75	68.95
Proposed CAMixer	1178	1203	2381	98.39	89.84

Method	Results on the Chao Lake II dataset				
	FP	FN	OE	PCC (%)	KC (%)
PCA-KM [16]	8432	2273	10705	92.74	65.69
NR-CR [17]	959	2397	3356	97.72	86.63
NR-ELM [18]	595	3836	4431	97.00	81.27
DDNet [19]	3107	779	3886	97.36	86.18
MSAPNet [20]	2006	837	2843	98.07	89.55
Proposed CAMiser	1416	1019	2435	98.35	90.84

C. Ablation Study

We conduct ablation experiments to verify the effectiveness of the PCAM and GFFN for the change detection task. We design the following four variants: 1) *Basic Network* represents the backbone without PCAM and GFFN. (2) *w/o PCAM* denotes the proposed method without PCAM, (3) *w/o GFFN* denotes the proposed method without GFFN, and (4) *w/o H-Clustering* denotes the proposed method employs fuzzy c-means for preclassification instead of hierarchical clustering [14].

The results in Table I demonstrate that compared to our full model, either *w/o PCAM* or *w/o GFFN* consistently exhibited lower performance on all datasets. This indicates that the PCAM significantly enhances the change detection performance, while the GFFN marginally improves it. It shows that GFFN enhances the non-linear feature transformation. Furthermore, the proposed method using hierarchical clustering demonstrates superior performance compared to *w/o H-Clustering*. It is apparent that hierarchical clustering generates more reliable training samples for the proposed CAMixer, consequently enhancing the change detection performance.

To further verify the validity of our proposed PCAM, we used the t-SNE [21] tool to visualize the characteristics before and after the module. As shown in Fig. 4, the feature representations after PCAM are noticeably more discriminative.

D. Experimental Results and Comparison

We compare the proposed CAMixer with five baselines, including PCA-KM [16], NR-CR[17], NR-ELM[18], DDNet [19] and MSAPNet [20]. Fig. 5 illustrates the visual comparison of the change maps generated by different methods on three datasets. The corresponding quantitative evaluation metrics are illustrated in Table II.

Results on the Yellow River dataset: The Yellow River dataset is severely degraded by speckle noise. As a result, it is difficult to obtain satisfactory results by conventional methods. The qualitative results of this dataset are shown in the first row of Fig. 5. It can be observed that the proposed CAMixer suppresses the false alarms effectively, and the change map of CAMixer is the most similar to the ground truth. Furthermore, the CAMixer reports the best PCC value, gaining 0.63% and 0.86% improvement of PCC over DDNet and MSAPNet, respectively. DDNet and MSAPNet are CNN-based methods, and it is evident that CAMixer improves the change detection performance by introducing the Transformer-like architecture.

Results on the Chao Lake I and II datasets: The qualitative results of Chao Lake I and II datasets are shown in the second and third rows of Fig. 5, respectively. The proposed CAMixer greatly reduces the false alarms, and obtains the best PCC and KC values on both datasets. It is evident that the proposed CAMixer improves the feature representations via parallel convolution and self-attention computation. The parallel design of shift convolution and self-attention extracts local and global features simultaneously, leading to high-quality representations against the speckle noise. Additionally, the GFFN selectively emphasizes critical features, which further mitigates the interference caused by speckle noise.

From the above experiments on three SAR datasets, it can be seen that the proposed CAMixer has better performance than several traditional methods and CNN-based methods. Furthermore, CAMixer reports the best KC values, gaining 1.90%, 12.24%, and 1.29% improvement over the second-best one on three datasets, respectively. It should be noted that KC value is of the most convincing evaluation metrics for SAR change detection. Moreover, the CAMixer obtains balanced FP and FN values on three datasets. It demonstrates that the PCAM captures abundant convolution and self-attention feature interactions, and contributes to better change detection results.

IV. CONCLUSIONS AND FUTURE WORK

In this letter, we propose CAMixer, a novel SAR change detection network that produces reliable change detection results. To address the inductive bias limitation of Transformer-like networks, we combine self-attention with shift convolution in a parallel manner. Moreover, we propose a gated feed-forward network to enhance non-linear feature transformation, formulated as the element-wise multiplication of two parallel linear layers. Extensive experiments on three SAR change detection datasets demonstrate the superiority of CAMixer and validate the effectiveness of its two critical components. In the future, we plan to investigate the fusion of mutli-source remote sensing data to improve the change detection performance.

REFERENCES

[1] J. G. Vinholi, B. G. Palm, D. Silva, R. Machado, and M. I. Pettersson, "Change detection based on convolutional neural networks using stacks of wavelength-resolution synthetic aperture radar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[2] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[3] M. Liu, X. Liu, L. Wu, T. Peng, Q. Zhang, X. Zou, L. Tian, and X. Wang, "Hybrid spatiotemporal graph convolutional network for detecting landscape pattern evolution from long-term remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[4] L. Nava, O. Monserrat, and F. Catani, "Improving landslide detection on SAR data through deep learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[5] H. Feng, L. Zhang, J. Dong, S. Li, Q. Zhao, J. Luo, and M. Liao, "Mapping the 2021 October flood event in the subsiding Taiyuan basin by multitemporal SAR data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7515–7524, 2022.

[6] D. Meng, F. Gao, J. Dong, Q. Du, and H.-C. Li, "Synthetic aperture radar image change detection via layer attention-based noise-tolerant network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[7] F. Liu, L. Jiao, X. Tang, S. Yang, W. Ma, and B. Hou, "Local restricted convolutional neural network for change detection in polarimetric SAR images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 818–833, 2019.

[8] S. Saha, M. Shahzad, P. Ebel, and X. X. Zhu, "Supervised change detection using prechange optical-SAR and postchange SAR data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8170–8178, 2022.

[9] J. Wang, F. Gao, J. Dong, Q. Du, and H.-C. Li, "Change detection from synthetic aperture radar images via dual path denoising network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2667–2680, 2022.

[10] S. Hafner, A. Nascetti, H. Azizpour, and Y. Ban, "Sentinel-1 and Sentinel-2 data fusion for urban change detection using a dual stream U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[11] Z.-G. Liu, Z.-W. Zhang, Q. Pan, and L.-B. Ning, "Unsupervised change detection from heterogeneous data based on image translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2021, pp. 1–21.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10 002, 2021.

[14] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on pcanet," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1792–1796, 2016.

[15] G. Wang, Y. Zhao, C. Tang, C. Luo, and W. Zeng, "When shift operation meets vision transformer: An extremely simple alternative to attention mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2423–2430.

[16] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 772–776, 2009.

[17] Y. Gao, F. Gao, J. Dong, and S. Wang, "Sea ice change detection in SAR images based on collaborative representation," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018, pp. 7320–7323.

[18] F. Gao, J. Dong, B. Li, Q. Xu, and C. Xie, "Change detection from synthetic aperture radar images based on neighborhood-based ratio and extreme learning machine," *Journal of Applied Remote Sensing*, vol. 10, no. 4, p. 046019, 2016.

[19] X. Qu, F. Gao, J. Dong, Q. Du, and H.-C. Li, "Change detection in synthetic aperture radar images using a dual-domain network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[20] R. Wang, F. Ding, J.-W. Chen, B. Liu, J. Zhang, and L. Jiao, "SAR image change detection method via a pyramid pooling convolutional neural network," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2020, pp. 312–315.

[21] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.