CHANGEMACS: A MAMBA-BASED ATTENTION GUIDED AND CONTRASTIVE SIMILARITY LEARNING NETWORK FOR REMOTE SENSING CHANGE DETECTION

*Kaixuan Jiang, Chen Wu**

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
*Corresponding author E-mail: chen.wu@whu.edu.cn

ABSTRACT

Recently, deep learning has made significant progress in modern remote sensing change detection (CD) with its powerful feature learning capabilities. However, most existing methods hardly capture multi-scale information in complex scenes. In addition, these methods lack the capability to model fine-grained feature variations, restricting the accurate localization of detail changes. To alleviate these issues, this paper presents a mamba-based attention guided and contrastive similarity learning network (ChangeMACS), ChangeMACS through mamba-based attention module (MBAM) to enhance the interaction ability between local and global features. Meanwhile, the contrastive similarity learning module (CSLM) is designed to dynamically reinforce the differences in changed regions and reduce the inconsistencies in unchanged regions based on the pixel-level similarity. The experimental results show that ChangeMACS has better performance in processing complex remote sensing images, which achieves State-of-The-Art (SoTA) results on SYSU-CD and LEVIR-CD datasets. Code is available at https://github.com/Jscript10/ChangeMACS.

***Index Terms*—** Change detection, Mamba-based attention, Contrastive similarity learning

1. INTRODUCTION

Change detection (CD) is an important research track in remote sensing, which is mainly used to identify the differences between imagery of the same geographic area acquired at different time points [1]. This technique has a wide range of applications in land use analysis, disaster assessment, and military surveys. With the rapid development of remote sensing imaging technology and high-resolution satellites, high-quality remote sensing data can provide richer surface information. However, these high-resolution data also bring unfavorable factors, such as pseudo-changes in shadows, noise, etc., which increase the complexity and challenge of CD tasks.

Traditional CD methods include algebraic operations, classification detection and image transformation approaches. Algebraic computing-based methods measure changes by calculating algebraic features between corresponding bands of multi-temporal remote sensing images. Classification-based methods mainly include post-classification comparison and joint classification, which are mainly used to obtain the change information and change types of features by means of classification. Image transformation-based methods mainly select bands and mathematically transform the image to recognize the main features of change. Typical methods contain wavelet transform (Wavelet Transform) and principal component analysis (PCA) [2]. However, traditional change detection methods have many limitations when dealing with complex scenes. Algebraic operation-based methods rely on simple computations between bands, have poor robustness to noise and pseudo-changes. Classification-based methods rely on high-quality training samples, and their results are easily limited by the performance of the classifier. Although image transformation-based methods can highlight the main change regions, they are more sensitive to noise and important detail information may be lost during the transformation process.

Deep learning-based CD methods have attracted significant focus in recent years, with remarkable advances. DSAMNet [3] introduces channels and spatial attention mechanism for better feature learning. DARNet [4] employs dense skip connections to capture long-range semantic features from potential space. Chen et al. [5] first brings Transformer to the CD task, which utilizes Transformer to capture global contextual information. SARASNet [6] combines CNN with Transformer to capture both local and global features. Although these approaches enhance the capabilities of CD networks, there are still some challenges. First, networks combining Transformer have high requirements for large-scale data and computational resources. Second, existing methods fail to adequately consider the difference and similarity relationship between bi-temporal images, which leads to the network's insufficient perception of changed regions.

To alleviate the above issues, we propose a novel method ChangeMACS, which can effectively enhance the network's attention to global information and accurate perception of

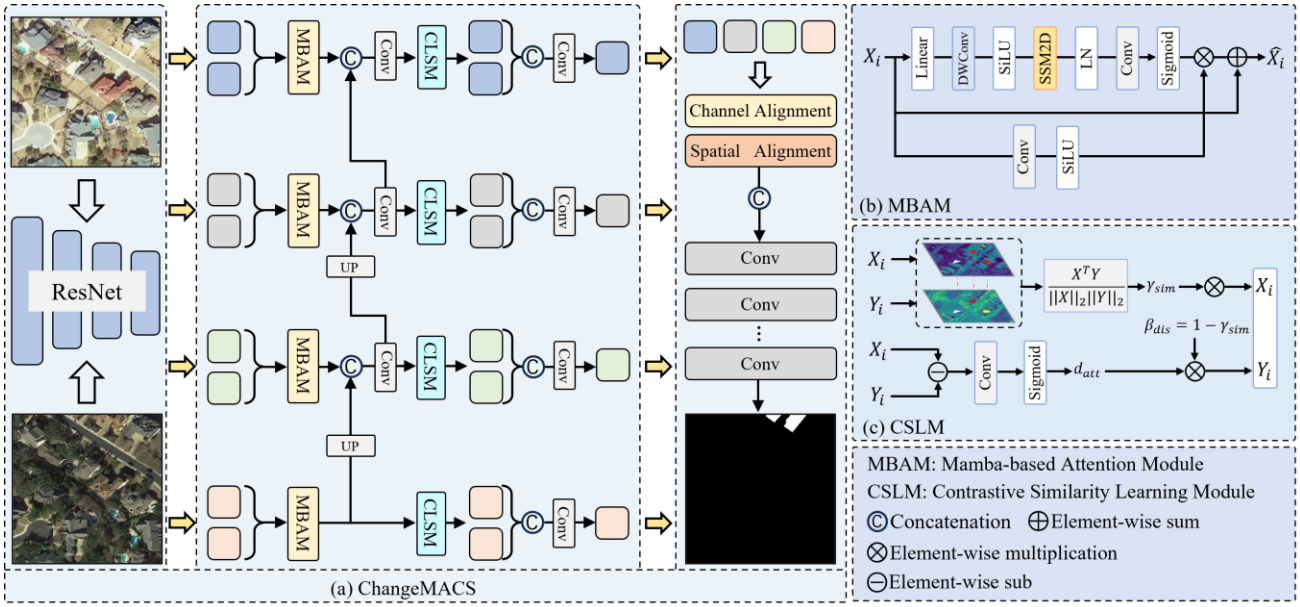

Fig.1 Overall architecture of ChangeMACS

changed regions. The main contributions of this paper are as follows:

- We constructed a new CD network that combines multi-scale feature extraction and fusion, which is able to accurately detect changed regions in complex scenes while maintaining a lightweight model.
- The mamba-based attention module (MBAM) is designed to enhance the interaction between local and global features, which enhances the modeling of fine-grained variations and long-range features.
- We propose the contrastive similarity learning module (CSLM), which dynamically reinforces the similarity and difference information of bi-temporal features through pixel-level cosine similarity computation and improves the network's ability to discriminate changed regions.

2. METHOD

2.1. Overview

The network structure of ChangeMACS is shown in Fig. 1(a), which contains two main components, Mamba-based Attention Module (MBAM) and Contrastive Similarity learning Module (CSLM). For the input bi-temporal image, it first undergoes feature extraction by ResNet [7] to obtain the initial features at different levels. Later, for each feature pair, we employ MBAM to enhance the global information modeling capability, which is used to capture the global contextual information. Then, a UNet-like structure is used to up-sample the four scales of features for fusion, and channel alignment is performed by convolution. After that, each layer of bi-temporal features is performed by CSLM for variation and invariance guidance, which enables the network to better distinguish between changed and unchanged regions. Then, the enhanced four scales of features and channels are concatenated and adjusted to the same size and channel by up-sampling and convolution operations. Finally, the four groups of features are concatenated and channel dimensionality reduction is performed by multilayer convolution to obtain the final detection results.

2.2. Mamba-based Attention Module

Our presented MBAM is shown in Fig. 1(b), which aims to enhance the global information modeling capability of the initial features. First, for the multi-scale features obtained from ResNet, we map the input features through a linear layer to reduce redundant information. After that, the features are processed by deep separable convolution [8] to fine-grain the spatial dimension. Subsequently, the features go to SSM2D [9], which uses chunking operations during processing to reduce the computational complexity of high-resolution remote sensing images by creating sparse relationships in localized blocks while preserving the global contextual information. Later, the weight coefficients of the features are generated by normalization and activation function, followed by element-wise multiplication of the weight coefficients with the input features, and joined by the residuals. This design allows the network to not only focus on local details, but also capture the global information in the image.

2.3. Contrastive Similarity Learning Module

The Contrastive Similarity Learning Module (CSLM) is proposed in Fig. 1(c), which aims to dynamically contrast and enhance the bi-temporal features to make the features of the similar regions more similar while significantly expanding the feature differences in the changed regions. The CSLM

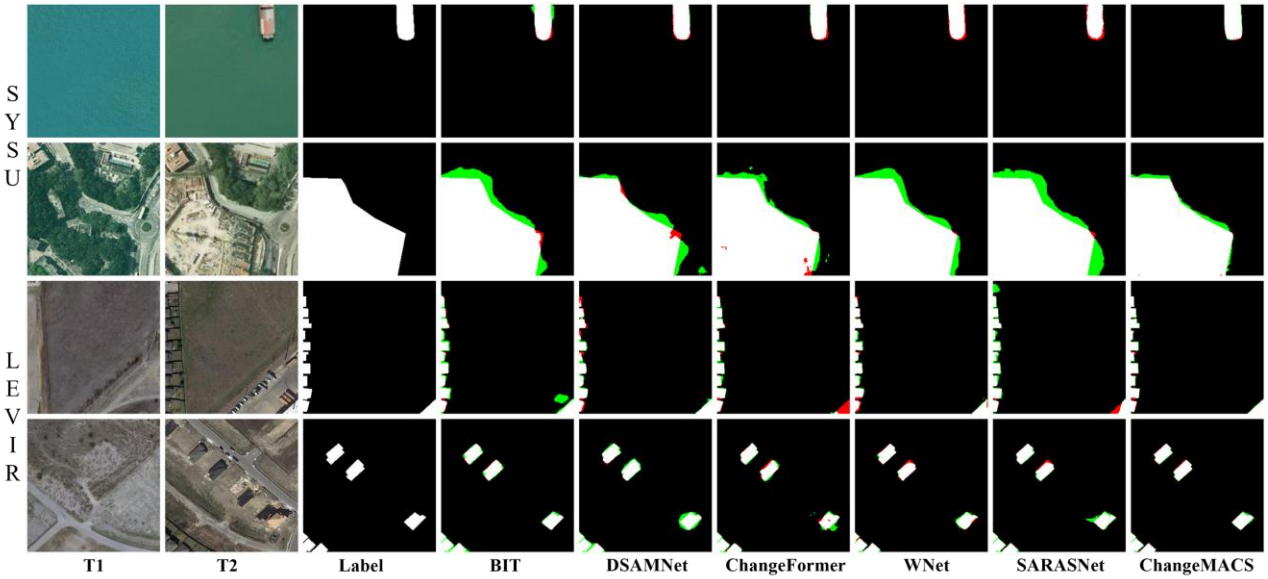

Fig.2 Qualitative results on the SYSU-CD (top two rows) and LEVIR-CD (bottom two rows) datasets.

Table 1. Quantitative results on SYSU-CD and LEVIR-CD datasets.

Method		SYSU-CD				LEVIR-CD			
	Params.(M)	Pre	Rec	IoU	F1	Pre	Rec	IoU	F1
BIT	11.47	77.53	80.63	65.35	79.05	87.38	91.92	81.15	89.60
DSAMNet	16.95	81.60	71.36	61.47	76.14	81.52	91.33	75.66	86.15
ChangeFormer	41.01	82.10	74.92	64.41	78.35	89.06	88.93	80.17	88.99
WNet	44.91	82.19	81.01	68.91	81.60	89.97	90.47	82.18	90.22
SARASNet	102.67	**84.34**	79.46	69.24	81.83	**90.23**	91.80	83.50	91.01
ChangeMACS	28.71	83.39	**83.32**	**71.45**	**83.35**	90.03	**92.67**	**84.05**	**91.33**

takes the bi-temporal features x and y as the inputs, and generates the difference features by calculating the element-wise absolute difference computation. To further separate the similar region from the changed region, the CSLM employs pixel-level cosine similarity computation to analyze the similarity of the bi-temporal features and generate similarity weights and disparity weights. Subsequently, the input features are multiplied element-wise with various weights. By accurately modeling similar and variation regions, CSLM provides more reliable feature representations for subsequent processing of CD network.

3. EXPERIMENTAL RESULTS

3.1. Dataset and Benchmark

In this paper, we conduct experiments on two popular CD datasets; SYSU-CD [3] and LEIVR-CD [10], SYSU-CD contains complex changes such as changes in ships, buildings and cities. We crop the images into 256×256 size to get 20,000 pairs of images, which are trained, validated and tested according to the ratio of 6:2:2. LEVIR-CD dataset mainly contains changes of buildings, with a total of more than 30,000 change objects, which covers a lot of cities in the U.S. To improve the training efficiency, we crop the images to 256×256 size. 7120 pairs of images for training, 1024 pairs of images for validation and 2048 pairs of images for testing. To validate the performance of ChangeMACS, we use five popular CD methods for comparison. These are BIT [5], DSAMNet [3], ChangeFormer [11], WNet [12] and SARASNet [6].

3.1. Experiment Settings

The experiments are performed by the pytorch platform, running on two NVIDIA RTX 3080 GPUs. we use the batch balanced loss function for loss calculation. AdamW is selected as the optimizer, learning rate is set to 1e-4, and the network is trained for 200 epochs. In terms of evaluation metrics, we focus on Recall (Rec), Precision (Pre), F1-score (F1) and Intersection over Union (IoU). Indicators are calculated as follows:

$$Pre = \frac{TP}{TP + FP} \tag{1}$$

$$Rec = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2Pre \cdot Rec}{Pre + Rec} \quad (3)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (4)$$

3.1. Comparison Experiments

Table 1 shows the quantitative analysis results of all approaches on SYSU-CD and LEVIR-CD datasets. On the SYSU-CD dataset, the Transformer-based methods, BIT and ChangeFormer achieved IoU of 65.35% and 64.41%, respectively. Among the CNN-based methods, DSAMNet and WNet achieved the IoU of 61.47% and 68.91%, respectively. the hybrid structure-based SARASNet achieved IoU of 69.24%. ChangeMACS achieved the best performance with IoU of 71.45% compared to the existing methods. In LEIVR-CD dataset, DSAMNet achieved the lowest IoU of 75.66%, BIT, ChangeFormer, WNet, and SARASNet achieved 81.15%, 80.17%, 82.18%, and 83.50% of IoU, respectively. Our method achieved the highest IoU with a score of 84.05%.

Fig. 2 shows the test results on two datasets, where red color indicates missed detection and green color indicates false detection. The test results on SYSU-CD are shown in the top two rows, where our method can capture the changed boundaries of small targets effectively. The test results on SYSU-CD are shown in the top two rows, where our method can capture the changed boundaries of small targets very effectively. Such as the ship change in the first column of graph, our method achieves the minimum false detection. For large targets, such as the second column, the compared methods all produce some degree of false detection, resulting in ambiguous detection edges. Our method can obtain detection results closest to the real changes, and the change boundaries are smoother. The results on LEVIR-CD are shown in bottom two rows. We can observe that for building changes with regular shapes, due to the integration of MBAM to enhance the local-global feature context information and CSLM to learn the variation and invariant features, our method achieves the best detection performance.

4. CONCLUSION

In this paper, we propose a mamba-based attention guided and contrastive similarity learning network (ChangeMACS). By introducing MBAM on the basis of local features extracted by ResNet, it enables the network to effectively enhance the local-global feature interactions to extract fine-grained semantic information. In addition, we enhance the feature consistency in similar regions through CSLM while magnifying the feature differences in changed regions, enabling the network to automatically focus on changed regions and suppress the interference of irrelevant backgrounds in the CD task. Experiments on two popular CD datasets show that ChangeMACS has high perception of changed regions and achieves SOTA performance. This indicates that our model can handle the change detection task in complex scenes with good robustness.

5. ACKNOWLEDGMENT

This work is partly supported by the National Natural Science Foundation of China under Grant T2122014, and partly by the National Key Research and Development Program of China under Grant 2022YFB3903300. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

6. REFERENCES

[1] Gong, M. G., L. Z. Su, H. Li, and J. Liu. "A survey on change detection in synthetic aperture radar imagery." Journal of Computer Research and Development 53, no. 1 (2016): 123-137.

[2] Deng, J. S., K. Wang, Y. H. Deng, and G. J. Qi. "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data." *International Journal of Remote Sensing* 29, no. 16 (2008): 4823-4838.

[3] Shi, Qian, 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* 60 (2021): 1-16.

[4] Li, Ziming, Chenxi Yan, Ying Sun, and Qinchuan Xin. "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-18.

[5] Chen, Hao, Zipeng Qi, and Zhenwei Shi. "Remote sensing image change detection with transformers." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021): 1-14.

[6] Chen, Chao-Peng, Jun-Wei Hsieh, Ping-Yang Chen, Yi-Kuan Hsieh, and Bor-Shiun Wang. "SARAS-net: scale and relation aware siamese network for change detection." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14187-14195. 2023.

[7] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[8] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251-1258. 2017.

[9] Wang, Chengkun, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. "V2M: Visual 2-Dimensional Mamba for Image Representation Learning." *arXiv preprint arXiv:2410.10382* (2024).

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

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

[12] Tang, Xu, 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* (2023).