Frequency-Temporal Attention Network for Remote Sensing Imagery Change Detection

Chunyan Yu, *Senior Member, IEEE*, Haobo Li, Yabin Hu, Qiang Zhang, *Member, IEEE*, Meiping Song, *Member, IEEE*, and Yulei Wang, *Member, IEEE*

***Abstract*— Change detection (CD) in remote sensing imagery is identified as a pivotal task in the field of Earth observation, while it usually confronts the dilemma of intricate data and minor alterations. To address the stated challenge, this letter presents an innovative frequency-temporal attention network for CD (FTAN), which incorporates two advanced modules including the multidimensional convolutional frequency attention module (MCFA) and the interactive attention module (IAM). Specifically, the MCFA module is essential for enhancing sensitivity in CD by merging multiscale spatial and frequency domain features. As a supplement to MCFA, the IAM aggregates category-related tokens and processes cross-attention information from different time phases. The seamless integration of MCFA and IAM empowers the FTAN network with enhanced capabilities to detect minor regions and edges accurately. Experiments on datasets like LEVIR-CD and DSIFN-CD demonstrate superior performance by outperforming existing models in F1 scores and IoU metrics. Our code and pretrained models will be released at https://github.com/chirsycy/FTAN.**

***Index Terms*— Adversarial training, domain adaptation, hyperspectral image (HSI) classification, transfer learning.**

I. Introduction

CHANGE detection (CD) [1] of remote sensing images (RSI) refers to automatically detecting differences or changes in multitemporal images of the same scene, which has great significance to the development of land cover, urban data collection, and environmental monitoring.

Generally speaking, RSI is characterized by nonlinear features including spatial and spectral variability. Traditional methods encounter challenges in dealing with the mentioned complexities. In recent years, deep learning (DL) methods [2], [3], [4], [5], [6], [7] have achieved great progress due to automatic feature extraction and superior performance in the fields of remote sensing applications. Nowadays, the attention-based model [8] for CD promotes the recognition of specific changed objects. In which, the self-attention mechanism, e.g., transformer-based network [8] captures long-range dependencies providing a more efficient way to encode spatial information. Typically, many models have employed spatial-temporal attention mechanisms to refine and improve features for CD implementations. TinyCD [9] combines low-level features for CD from global temporal and local spatial information for spatiotemporal feature fusion. ChangeFormer [10] directly extracts CD-related information from the input images and performs context modeling to achieve efficient and accurate detection results. Recently, some CD models have focused on addressing edge detection challenges. In particular, EGDE-Net [11] presents the edge-aware module for boundary information refinement. Changer [12] builds a new CD pattern with the interaction between bi-temporal features that is beneficial to details and edges. Recently, in the computer vision field, frequency domain learning has been a popular way to increase channel attention, which captures global patterns and long-range dependencies. Although the existing edge-based CD models have produced impressive results, the absence of frequency domain information results in incomplete or inaccurate detection of boundaries and minor changes. The RS-CD methods involved with the frequency domain effectively filter noise and enhance significant features for subtle changes.

In this letter, we proposed a novel frequency-temporal attention network for CD (FTAN), in which the multidimensional convolutional frequency attention module (MCFA) skillfully integrates multiscale spatial features with frequency domain characteristics. Notably, the frequency-domain masking-based adaptive convolution (FDM-AC) that is the core of MCFA excels at reducing noise and highlighting key signals, which is crucial in boosting the sensitivity of CD. Moreover, the interactive attention module (IAM) merges interactive spatial data by adaptive attention mechanism and enriches local and global feature representation, which is effective for region continuity and promotes detection connectivity. The primary contributions of this study are summarized as follows:

1) Distinguished from prior models, we propose a novel frequency-spatial-temporal attention network to enhance edge representation for CD of RSI. To the best of our knowledge, it is the first attempt to combine the frequency information in spatial-temporal extraction in the CD framework. Notably, the core MCFA mechanism integrates multiscale attention with frequency domain information to strengthen the spatial-temporal feature, which is beneficial for minor target detection.
2) Unlike the previous approaches, the presented IAM captures interactive long-range dependencies. As a critical part of the model, the IAM aggregates category-related

Received 26 June 2024; revised 9 September 2024; accepted 27 September 2024. Date of publication 10 October 2024; date of current version 24 October 2024. This work was supported in part by the National Nature Science Foundation of China under Grant 62471079 and Grant 62401095 and in part by the Fundamental Research Funds for the Central Universities under Grant 3132017124. *(Corresponding author: Qiang Zhang.)*

Chunyan Yu, Haobo Li, Qiang Zhang, Meiping Song, and Yulei Wang are with the Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: yucy@dlmu.edu.cn; lihaobo1998@gmail.com; qzhang95@dlmu.edu.cn; smping@163.com; wangyulei@dlmu.edu.cn).

Yabin Hu is with the First Institute of Oceanography and the Technology Innovation Center for Ocean Telemetry, Ministry of Natural Resources, Qingdao 266061, China (e-mail: huyabin@fio.org.cn).

Digital Object Identifier 10.1109/LGRS.2024.3477991

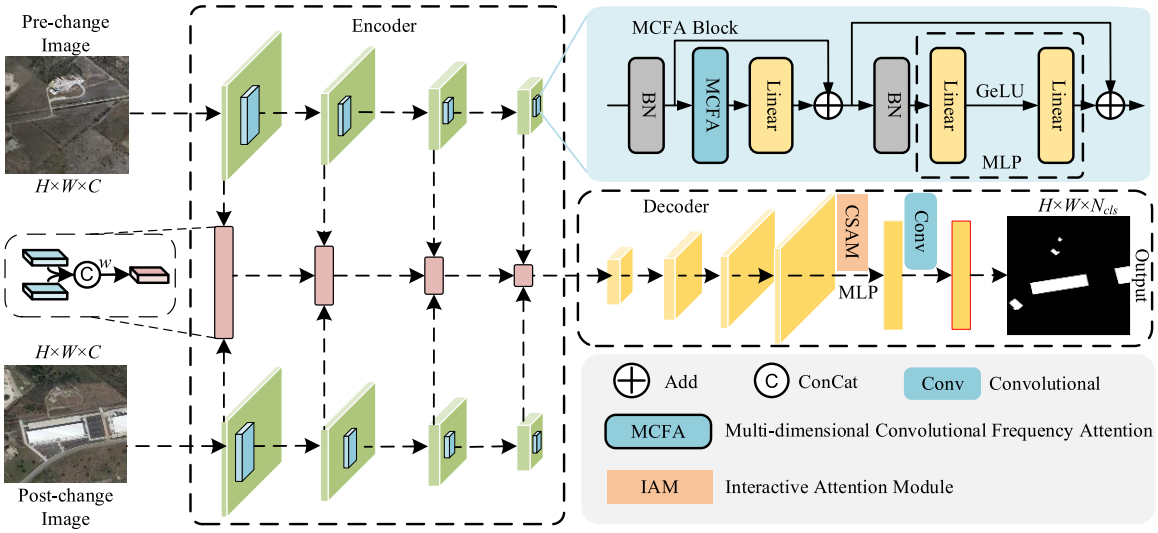


Fig. 1. Illustration of the proposed FTAN architecture.

tokens from multiple regions and adapts to the spatiotemporal context.

II. Proposed Approach

The overall architecture of the proposed FTAN is illustrated in Fig. 1. As observed, the encoder exploits novel frequency-perception convolution attention to extract saliency, and the decoder block with IAM is responsible for refining and generating the continuous detection map on partial regions. Further details are outlined in Sections II-A and II-B.

A. Multidimensional Convolutional Frequency Attention

1) Encoder Design: We employ a convolutional saliency extraction substituting for attention extraction. On the whole, the encoder of our approach contains four stages for refined saliency extraction with the decreasing resolution of $H/4 \times W/4$, $H/8 \times W/8$, $H/16 \times W/16$, $H/32 \times W/32$, where H and W represent the height and width of the input image, respectively. The pattern maintains the salient feature of the minor changes with the self-attention supplied by the convolutional saliency.

2) MCFA: As shown in Fig. 2, the primary elements of MCFA consist of FDM-AC, partial convolution operation (PConv), and multibranch depth-wise strip convolution block. Specifically, four parallel blocks are built in FDM-AC. In each block, a spatial domain image X is first converted into the frequency domain representation via the fast Fourier transform (FFT), as defined in the subsequent equation

$$X_{\text{freq}} = \mathcal{F}(X). \tag{1}$$

Afterward, a frequency mask scaled by a predefined parameter is created to match the size of the X_{freq}, and the mask is defined as follows:

$$\text{mask} = \begin{cases} 1, & \left|u - \frac{M}{2}\right| < \text{scale and } \left|v - \frac{N}{2}\right| < \text{scale} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where M and N denote the row and column and u and v represent the horizontal and vertical coordinates of the frequency

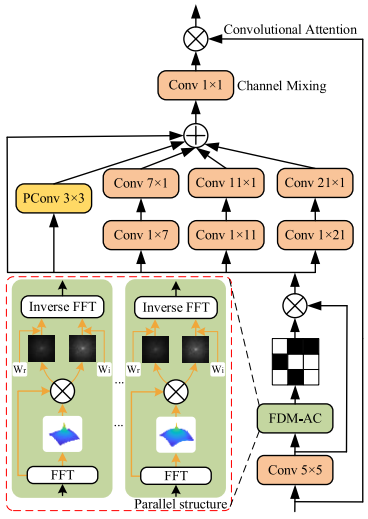


Fig. 2. Illustration of the MCFA. Primarily, FDM-AC generates confused features with frequency domain that is shown in the parallel structure, PConv preserves information from the remaining channels, and multibranch depth-wise strip convolution block employs $1 \times n$, $n \times 1$ kernels for spatial detail capture.

domain. With the mask, we obtain a filtered frequency domain image of RSI denoted as X_{masked} by the following equation:

$$X_{\text{masked}} = X_{\text{freq}} \cdot \text{mask}. \tag{3}$$

According to the inherent properties of the FFT, X_{masked} is decomposed of the real X_{real} and imaginary X_{imag} parts. To further mine the frequency domain information of the RSI, two convolutional operations with different weights denoted as w_r and w_i are employed to process the real and imaginary parts, respectively. Notably, $w_r, w_i \in C^{1\times c\times 1\times 1}$, where C denotes the set of complex numbers and c is the number of channels.

With the fusion operation, the output embedding $\mathbf{X}'$ of FDM-AC is yielded by (4), where $*$ means the convolution operation and $\mathcal{F}^{-1}$ represents the inverse FFT

$$X' = \left| \mathcal{F}^{-1} \frac{1}{4} \sum_{i=1}^{4} \left(X_{\text{real}} * w_r + i \left(X_{\text{imag}} * w_i \right) \right) \right|. \tag{4}$$

In MCFA, Pconv is adopted to preserve saliency information to improve boundary and detail prediction due to the convolution kernels only on a subset of input tensor channels. The implementation formula for Pconv is shown as follows:

$$W_p(X_f) = \text{Concat}(W_{3\times3}(X_1), X_2) \tag{5}$$

where X_f represents $X \otimes X'$ and $\otimes$ is a matrix multiplication operator and $W_{3\times3}$ denotes the 3×3 convolutional operation. X_1 denotes the preceding portion of channels of X_f, and X_2 refers to the remaining channels of X_f.

Additionally, the deep-wise strip convolution in RSI utilizes $1 \times n$, $n \times 1$ kernels for spatial detail capture in changed objects. Specifically, the saliency information extracted from MCFA is obtained with the following formula:

$$X_h = W_i(X_f) \tag{6}$$

$$\tilde{X}_h = X_h \oplus W_p(X_h) \oplus \left\{W_n'(W_n(X_h))\right\} \tag{7}$$

where W_i denotes the convolutional operation with a kernel size of 5×5, X_h denotes the output of the W_i applied to X_f and $\tilde{X}_h$ is the concatenated maps with multiple branches. $\oplus$ is the concatenation operation and W_p represents Pconv. Besides, W_n and W_n' represent the convolutional operation with the kernel size of $1 \times n$ and $n \times 1$, $n \in \{7, 11, 21\}$, respectively.

Last, the refined feature of X is achieved via an attention mechanism that is guided by a 1×1 convolutional layer

$$H_{\text{out}} = W_o(\tilde{X}_h) \otimes X \tag{8}$$

where H_{out} represents the output of MCFA, $\otimes$ is a matrix multiplication operator, and W_o denotes the 1×1 operation.

B. Interactive Attention Module

As shown in Fig. 3, the IAM module is located at the end of the decoder of the FTAN. Structurally, the decoder employs a cascade of four upsampling stages to progressively increase the spatial resolution, with the incorporated residual blocks serving to refine the features. Specifically, the token sequences are composed of Inter tokens and Cls tokens, where the Inter tokens are obtained through a linear transformation of the feature maps, and the Cls token is a category-related token that is randomly initialized during the initialization process. Eventually, the decoder concludes by processing the refined feature maps via a 3×3 convolutional layer. Notably, the Cls token is employed as an interactive bridge and effectively integrates and compares data from different temporal images. Besides, it only serves as keys and values in the self-attention computation, which enables the model to directly access and process spatiotemporal data, facilitating the handling of complex spatiotemporal information. The self-context attention is calculated as follows:

$$H_t = \underset{T}{\text{Softmax}}\left(q(s_t) \times k(s_t)^T \cdot C^{-\frac{1}{2}}\right) \cdot v(s_t) \tag{9}$$

where s_t is the input from inter tokens, t denotes the pixel position, q is the linear mapping from s_t to the query, k denotes the linear mapping from s_t to the key, v denotes the linear mapping from s_t to the value, T is the total number of pixels, C is the number of channels of the input feature map s_t, and H_t denotes the intermediate feature by self-attention mechanism.

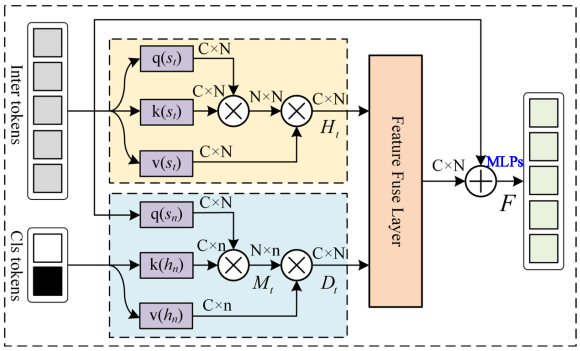


Fig. 3. Illustration of the IAM. First, Inter tokens and CLS tokens are obtained. Next, self-context and interactive spatiotemporal information enhancement are implemented by (9)–(12). Subsequently, the feature fusion layer is responsible for feature aggregation and acquires a map for the following change prediction.

Specifically, the computation of cross-spatiotemporal information enhancement is calculated as follows:

$$M_{t,n} = \underset{N}{\text{Softmax}}\left(\frac{q(s_n) \times k(h_n)^T}{\sqrt{C}}\right) \tag{10}$$

$$Z_t = \sum_{n=1}^{N} M_{t,n} \tag{11}$$

$$D_t = \frac{1}{Z_t}\sum_{n=1} M_{t,n} \odot v(h_n) \tag{12}$$

where n denotes the feature pixel position. h_n represents the input feature from cls tokens, k represents the linear mapping from h_n to the key, and v denotes the linear mapping from h_n to the value. $M_{t,n}$ represents the similarity corresponding to the pixel at the position t. Z_t denotes the aggregate of n pairwise similarities computed using the pixel t and $\odot$ is the Hadamard product of the matrix. D_t denotes the intermediate feature by the cross-attention mechanism. After being processed by the feature fusion layer, the outputs of D_t and H_t are added with the input Inter tokens to constitute the final IAM feature map denoted as F.

Subsequently, F is initially processed through an MLP layer and unsampled to the size of $H \times W$. Finally, the obtained feature maps are processed through another MLP layer to predict the change mask with a resolution of $H \times W \times N_{\text{cls}}$. Usually, N_{cls} is set to 2 in the CD task.

III. Experiment and Result Analysis

A. Data Description

In this section, we evaluate the proposed FTAN with two popular RSI datasets.

1) LEVIR-CD: The dataset comprises 637 pairs that cover diverse building types with a resolution of 1024×1024. We acquired samples with a size of 256×256 for model training by cropping the original image. Separately, the training set, the validation set, and the testing set contain 7120, 1024, and 2048 samples in the following experiments.

2) DSIFN-CD: The dataset involves six categories of satellite RSI pairs gathered by urban areas in China, and the original resolution is 512×512. Similarly, we cropped the image into samples with 256×256 resolution, and the training set, validation set, and testing set include 14 400, 1360, and 192 pairs, respectively.

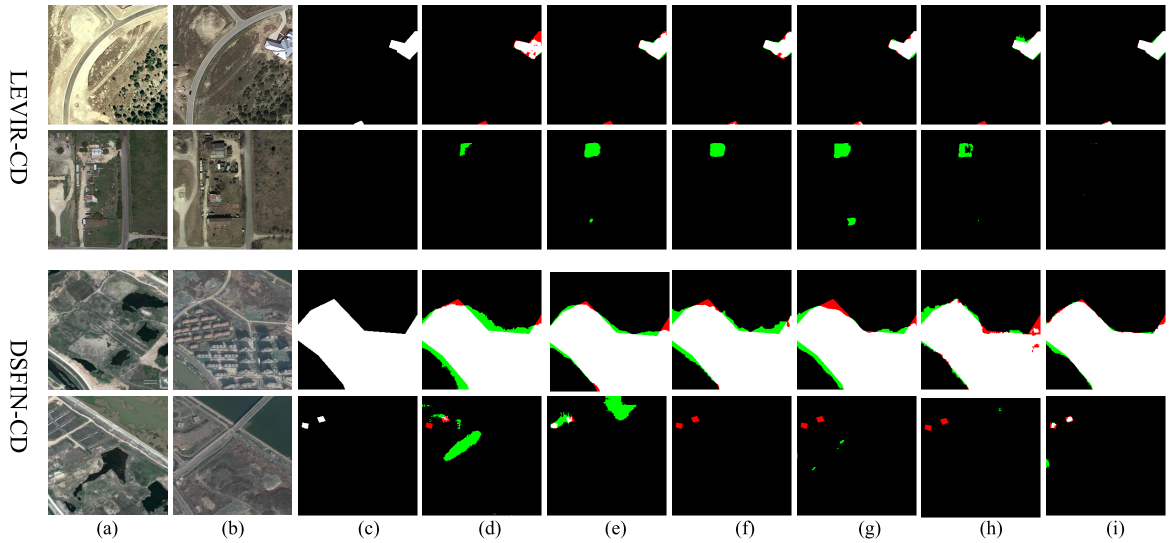


Fig. 4. Comparison results of different CD methods on LEVIR-CD (the first two rows) and DSIFN-CD (the last two rows). (a) Pre-Img. (b) Post-Img. (c) Label. (d) FC-EF. (e) DTCDSCN. (f) BIT. (g) TinyCD. (h) ChangeFormer. (i) FTAN. (Green indicates false detection, while red denotes missed detection).

B. Experimental Settings

All experiments were conducted on a computer in the platform of PyTorch with an NVIDIA Quadro RTX 8000 GPU for training. In the training phase, we perform data enhancement by random flipping, rescaling operation adjusts the image size by a factor randomly chosen between 0.8 and 1.2, cropping, gaussian blurring, and random color dithering to increase the number of training samples. Besides, the cross-entropy loss and AdamW optimizer are adopted in the experimental execution. Besides, the learning rate is initially set to 0. 0001, and the batch size is fixed at 24. In particular, all the CD approaches have no pretrained model for fair comparison in the following experiments.

C. Results and Analysis

First, we verify the performance to demonstrate the superiority of FTAN. First, we exploited a series of SOTA methods for comparison with our proposed approach. Typically, TinyCD adopts Siamese U-Net architecture that employs low-level features to achieve efficient CD. The ChangeFormer model builds a hierarchical Transformer encoder and MLP decoder to yield the detection map. The experiment results with all the CD approaches on LEVIR-CD and DSIFN-CD are reported in Table I and Fig. 4. As can be observed, FTAN generates the highest F1 score and IoU. Specifically, the F1 score and IoU on the LEVIR-CD dataset are 90.51% and 82.78%, respectively. While for the DSIFN-CD dataset, the values are 89.56% and 81.10%, respectively. As shown in Fig. 4, both the green regions and the red regions of the FTAN are the fewest, which indicates false and missed detection rates are the lowest. Compared with the current state-of-the-art ChangeFormer model, the proposed new CD paradigm is beneficial for the sensitivity to the changed targets with different scales and mottled regions. All the results and analyses demonstrate the effectiveness and stability of our model.

Next, the ablation study is performed to verify the contribution of the two modules in the FTAN. Individually, Tables I and II exhibit the ablation results. As demonstrated in Table I, regarding the LEVIR-CD dataset, the presented model that employed both MCFA and IAM yields the best performance such as Precision, Recall, F1, and IoU. The implementation with IAM is more competitive than the model without it, and the precisions are 91.54% and 92.12%, respectively. Although the implementation with either MCFA generates an improvement in detection, the approach leads to an improvement in detection, and the approach equipped with both MCFA and IAM yields the best performance on the four criteria. Specifically, Precision, Recall, F1, and IoU values reach 92.41%, 88.82%, 90.51%, and 82.78%, respectively.

For the DSIFN-CD dataset, the ablation study reveals a similar trend, where the fusion of MCFA and IAM modules significantly augment performance metrics. As detailed in Table II, the dual-module configuration achieves enhancing Recall and IoU to 89.56% and 81.10%, respectively. The results robustly validate that the combination of MCFA and IAM not only amplifies the detection ability to subtle variances within the dataset but also appreciably reduces false positive rates. Briefly, as observed, for the LEVIR-CD dataset, the MCFA module provides greater benefits in terms of precision and IoU. For the two datasets, all evaluation metrics are enhanced by the inclusion of the IAM modules as shown in the first and third rows of Tables II and III.

Moreover, we further undertook a study to evaluate the efficacy of the FDM-AC within the MCFA component. Fig. 5 demonstrates the generated activation maps via the Grad-CAM algorithm [16]. As can be observed, the visual maps in Fig. 5(d) indicate the augmented detection precision facilitated by FDM-AC, which aligns activation maps with predictive outcomes. Compared to the model without FDM-AC, the approach with FDM-AC yields a more interesting area of CD and strengthens the difference between targets and noise. As depicted in the figure, by effectively amplifying feature representation, the FDM-AC empowers the MCFA to recognize the subtle distinctions with multiple information involved, which improves accuracy for CD of RSI.

TABLE I
AVERAGE QUANTITATIVE RESULTS WITH DIFFERENT CD METHODS ON THE TWO DATA SETS

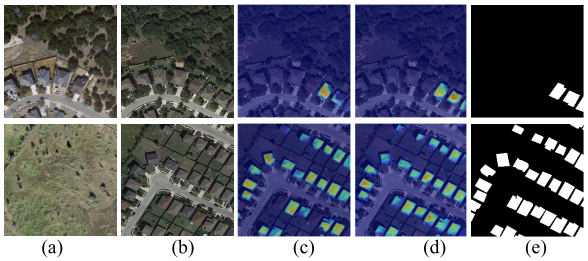
Model	LEVIR- CD				DSIFN-CD			
	Precision	Recall	F1	IoU	Precision	Recall	F1	IoU
FC-EF[13]	89.31	84.81	87.01	77.00	70.16	68.45	69.30	53.02
FC-Siam-Diff[13]	87.89	75.91	81.46	68.72	72.26	49.56	58.80	41.64
FC-Siam-Conc[13]	85.93	80.02	82.87	70.75	55.18	72.95	62.83	45.80
DTCDSCN [14]	90.35	85.82	88.03	78.62	80.37	83.88	82.09	69.62
BIT[15]	90.61	88.25	89.41	80.86	81.83	75.88	78.74	64.94
ChangeFormer[10]	91.54	87.04	89.23	80.56	89.40	85.93	87.63	77.98
TinyCD[9]	90.74	**89.03**	89.88	81.62	77.48	77.60	77.54	63.32
FTAN (ours)	**92.41**	88.82	**90.51**	**82.78**	**90.54**	**88.61**	**89.56**	**81.10**

TABLE II
ABLATION STUDY ON LEVIR-CD

MCFA	IAM	Precision	Recall	F1	IoU
✗	✗	91.54	87.04	89.23	80.56
✓	✗	92.12	87.70	89.85	81.58
✗	✓	91.88	88.51	90.17	82.09
✓	✓	**92.41**	**88.82**	**90.51**	**82.78**

TABLE III
ABLATION STUDY ON DSIFN-CD

MCFA	IAM	Precision	Recall	F1	IoU
✗	✗	89.40	85.93	87.63	77.98
✓	✗	88.79	**89.01**	88.99	80.17
✗	✓	89.60	87.92	88.75	79.78
✓	✓	**90.54**	88.61	**89.56**	**81.10**



(a) (b) (c) (d) (e)

Fig. 5. Grad-CAM Visualization of MCFA Efficacy. (a) Pre-Img. (b) Post-Img. (c) Grad-CAM Activation without FDM-AC (d) Grad-CAM Activation with FDM-AC. (e) Label.

IV. CONCLUSION

This letter presents a CD model that leverages the collaboration of spatial features, frequency information, and interchannel saliency. The proposed FTAN network incorporates the MCFA model and the IAM block, which effectively merges multiscale spatial features with frequency domain characteristics to enhance CD precision. Notably, MCFA focuses on extracting features through multiscale convolutional attention. Additionally, the integration of IAM in the decoder of the FTAN improves feature discrimination by efficiently aggregating category-related temporal data. Extensive experiments and analysis demonstrate superior performance compared to existing advanced models. In the future, we plan to combine multisource RSI to fully exploit the complementary information and enhance CD accuracy further.

REFERENCES

[1] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[2] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[3] C. Yu, Y. Zhu, M. Song, Y. Wang, and Q. Zhang, "Unseen feature extraction: Spatial mapping expansion with spectral compression network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5521915.

[4] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[5] C. Yu, M. Xu, Q. Zhang, and X. Lu, "Dual-intervention-constrained mask-adversary framework for unsupervised domain adaptation of hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[6] H. Yu, H. Yang, L. Gao, J. Hu, A. Plaza, and B. Zhang, "Hyperspectral image change detection based on gated spectral–spatial–temporal attention network with spectral similarity filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[7] Y. Wang, X. Chen, E. Zhao, C. Zhao, M. Song, and C. Yu, "An unsupervised momentum contrastive learning based transformer network for hyperspectral target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 9053–9068, 2024.

[8] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[9] A. Codegoni, G. Lombardi, and A. Ferrari, "TINYCD: A (not so) deep learning model for change detection," *Neural Comput. Appl.*, vol. 35, no. 11, pp. 8471–8486, Apr. 2023.

[10] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 207–210.

[11] Z. Chen et al., "EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 203–222, Sep. 2022.

[12] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610111.

[13] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[14] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[15] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.