Activated-SAM for Remote Sensing Change Detection Network

Wei Wang[1], Aocheng Shu[2], Xin Wang[1]
[1]School of Physics and Electronic Science, Changsha University of Science and Technology, Changsha, China
[2]School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China

***Abstract*—Due to the complex land cover types typically present in remote sensing images, it is difficult to achieve more precise detection of changed areas. Inspired by the Segment Anything Model (SAM), a change detection network named ASFCD is proposed incorporating an activated SAM, which utilizes a novel SAM encoder that incorporates NewActNet layers in place of the original MLPs. This encoder, termed Activated-SAM, enables dynamic feature filtering with a 44.54% reduction in parameter count compared to the original SAM encoder. Specifically, a dual-encoder architecture integrating Activated-SAM and VGG is proposed to extract features at different resolutions. Through the SAM-assisted Global-Local Fusion Module (SAGL), both dual-encoder feature fusion and bitemporal feature fusion, finally generate the output layer by layer. ASFCD was extensively evaluated on the public change detection datasets LEVIR-CD and WHU-CD, and its generalization capability was further validated on the SYSU-CD dataset. On the LEVIR-CD dataset, ASFCD achieved an F1-score of 92.18% and an IoU of 85.50%, outperforming the second-best model by 0.42% and 0.72%. On the WHU-CD dataset, it attained an F1-score of 94.84% and an IoU of 90.18%, exceeding the second-best model by 1.58% and 2.8%, achieving state-of-the-art performance. Our code can be seen at https://github.com/zcbg123/ASFCD.git.**

***Index Terms*—RSCD, Activated, SAM**

I. Introduction

Remote Sensing Change Detection (RSCD) is a process designed to identify changed regions in multi-temporal remote sensing images. With the increasing accessibility of high-precision remote sensing imagery, change detection has demonstrated growing value across numerous domains, such as global resource surveillance [1] and land resource monitoring [2]. Numerous existing change detection methods are based on convolutional neural networks (CNNs) [3] [4] [5]. Compared to traditional approaches, CNN-based deep learning significantly reduces the reliance on manual intervention and lowers the computational workload. However, CNNs remain ineffective in processing global information. As shown in Fig. 1, methods relying solely on CNN, while achieving relatively satisfactory results in local detection, are prone to false positives and missed detections when handling global information.

Recently, the emergence of visual foundation models (VFMs) has offered a fresh perspective for computer vision research. Among these, the Segment Anything Model (SAM) [6] has demonstrated the advantages of large-scale models in domains such as image segmentation [7] [8] [9], owing to its strong generalization capabilities. Precisely, SAM is implemented based on the Vision Transformer (ViT) [10], which effectively captures global contextual information, thereby enabling accurate segmentation of complex scenes. This capability potentially complements the limitations of CNNs. However, directly applying SAM to remote sensing change detection poses significant challenges due to the substantial domain gap (e.g., pseudo-changes) between remote sensing images and natural images, as well as the differences in objectives between change detection and segmentation tasks. Furthermore, the high computational resources required by the SAM model make its practical application a major focus of current research.

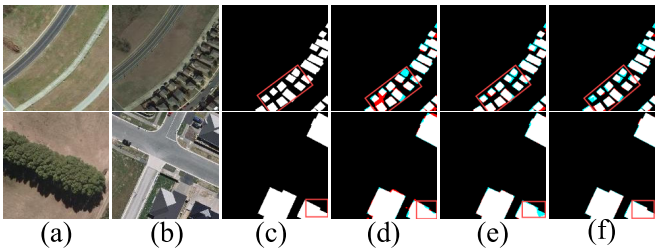

Fig. 1. Visualization results of different change detection methods. (a)T1 image. (b)T2 image. (c)Label. (d)SEIFNet [14]. (e)SFEARNet [15]. (f)ASFCD. TP, TN, FN, and FP are represented by white, black, blue, and red, respectively.

Additionally, research on alternatives to Multilayer Perceptrons (MLPs)that struggle with long sequences has become a major research focus, with examples including Kolmogorov-Arnold Networks (KANs) [11] [12] and ActNet [13]. While KANs address the issue of excessive parameter numbers, they impose stringent hardware requirements and still incur substantial training time costs; however, ActNet mitigates the training cost problem of KANs, though its broad applicability across deep learning remains uncertain.

To address the aforementioned issues, an auxiliary network adapted for remote sensing change detection named ASFCD is proposed, which employs NewActNet to replace the MLP layers in the SAM encoder, thus transforming it into an "activated" Activated-SAM. The features extracted by the Activated-SAM and those from the traditional VGG encoder are processed through the SAM-assisted Global-Local Fusion Module to handle both dual-encoder features and bitemporal features, ultimately generating the final change map layer by

Corresponding author: Xin Wang (wangxin@csust.edu.cn)

layer. Our contributions can be summarized as follows:

(1) To achieve both lower computational costs than MLPs and effective processing of long-sequence information, the Activated-SAM module is proposed. This module primarily consists of a SAM encoder wherein the original MLP layers have been replaced with NewActNet, thereby leveraging the combined advantages of both SAM and ActNet.

(2) To achieve effective integration of local and global information, a SAM-assisted Global-Local Fusion Module (SAGL) is proposed. This module performs dual-encoder feature fusion and bitemporal feature fusion on multi-level features extracted by both VGG and Activated-SAM, followed by local-global attention processing.

(3) Extensive comparative and ablation experiments were conducted on two public remote sensing change detection datasets to validate its detection capability. Its generalization performance was also rigorously evaluated. Experimental results demonstrate that ASFCD achieves highly competitive performance.

II. METHOD

A. Overall Structure

Fig. 2 illustrates the overall network architecture of ASFCD. Bitemporal images with dimensions H×W×3 are fed into the main encoder (VGG-16) and the auxiliary encoder (Activated-SAM), respectively. The H×W images are progressively downsampled to H/16×W/16 resolution, with channel dimensions successively increasing to 128, 256, 512, and 512, thereby generating dual-encoder features. Features generated by the auxiliary encoder are directly passed to the next stage, whereas features $V_{ti}^{k}(\ i \in \{1,2\}, k \in \{1,2,3,4\})$ extracted by the main encoder undergo hierarchical interactive fusion to yield VP_{ti}^{k}, thereby complementing feature information across different levels. Taking the first-layer feature V_{ti}^{1} with 128 channels from a certain temporal phase as an example, features with other channel dimensions will be aligned to 128 channels via convolutional operations. This process can be expressed as:

$$CA_1\left(x^{h\times w\times 128}\right) = Conv_{3\times 3}\left(x^{h\times w\times c}\right) \tag{1}$$

$$VP_{Ti}^{1} = Cat\left(V_{ti}^{1}, CA_1\left(V_{ti}^{2}\right), CA_1\left(V_{ti}^{3}\right), CA_1\left(V_{ti}^{4}\right)\right) \tag{2}$$

where CA_1 denotes the channel-aligned first-layer features.

To better handle long-range global information and enable global-local interaction, the hierarchically interactive features from the main encoder and the features generated by the auxiliary encoder are fed into the SAGL Fusion Module. This module performs dual-encoder feature fusion and bitemporal feature fusion, resulting in dual-encoder bitemporal fused features. The fused features then undergo local-global attention processing, followed by layer-by-layer output of the change map.

B. Activated-SAM

The fully connected layers used in MLPs cause the parameter count to grow quadratically with increasing input dimensions and network depth. This not only leads to higher computational costs in Transformer-based architectures but also results in limited capability for processing long sequences. To address these issues, a NewActNet(NA) module was designed to be compatible with Transformer architectures as a replacement for MLPs, as illustrated in Fig. 3. This module effectively reduces the high computational overhead associated with MLPs while possessing long-sequence processing capability comparable to that of KANs. This module employs two nested ActLayer and LayerNorm layers, where the NewActLayer can be regarded as a multi-head MLP layer. Each head first applies its individual activation function ϕ_i , followed by a linear layer $\lambda_i = \gamma_i$ with an output dimension of 1. Expressing the entire process in matrix form, let $\mathbf{\Phi}_{ij}=\phi_i(x_j)$ and $\mathbf{\Lambda}_{ij}=\lambda_{ij}$, and define $S:\mathbb{R}^{a\times b}\to\mathbb{R}^{a}$ as the function returning row sums of a matrix. Thus, the operational process of the entire NA can be expressed as follows:

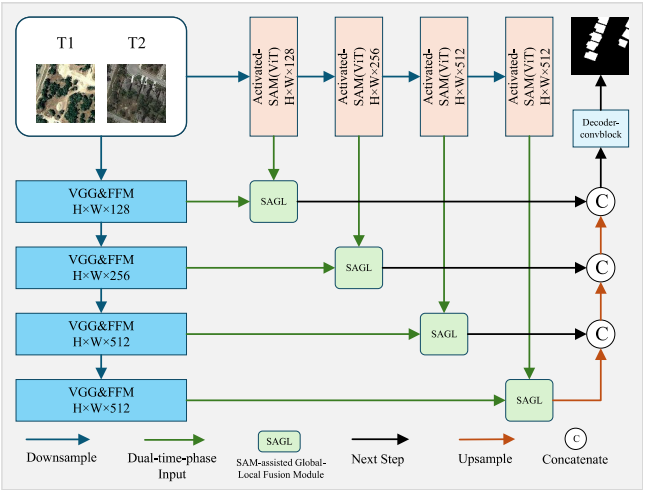

Fig. 2. The architectural framework of ASFCD.

$$NewActLayer_{\mathbf{\Lambda}}(x) = S(\mathbf{\Lambda}\odot\mathbf{\Phi}(x)) \tag{3}$$

$$NA(x) = LN(NewActLayer(LN(NewActLayer(x)))) \tag{4}$$

where $\odot$ denotes the Hadamard product, LN denotes LayerNorm, which functions similarly to the Norm layers and NA denotes NewActNet.

At this point, the ViT within the Activated-SAM (A-S) is activated into an Activated-ViT, while its other components remain structurally nearly identical to the original SAM encoder. To fully adapt its pre-trained representations to the remote sensing change detection task, all parameters of the Activated-SAM encoder are fine-tuned end-to-end during training. To ensure seamless fusion with the features extracted by the main encoder, the feature dimensions extracted by A-S are maintained consistent with those of the main encoder. Suppose A-S extracts features $S_{ti}^{k} \in \mathbb{R}^{C\times H\times W}$ and obtains AS_{Ti}^{k}. The entire process can be formally expressed as follows:

$$AS_{Ti}^{k} = NA\left(MSA\left(LN\left(\left(\ S_{ti}^{k}\right)\right)\right)\right) + S_{ti}^{k} \tag{5}$$

where MSA denotes Multi-head Self-Attention.

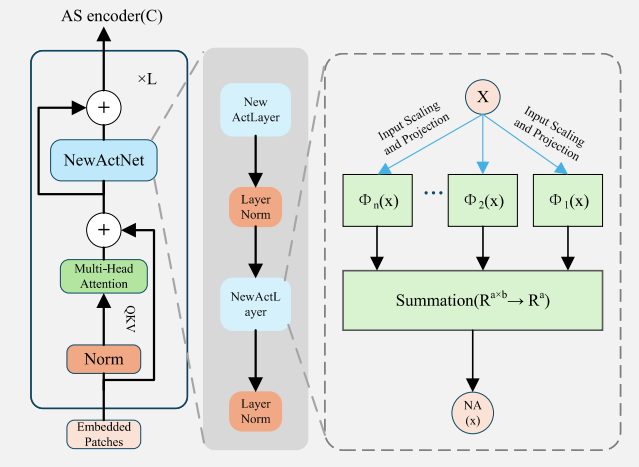

Fig. 3. The Transformer architecture within Activated-SAM, where C denotes the channel dimension of the extracted features.

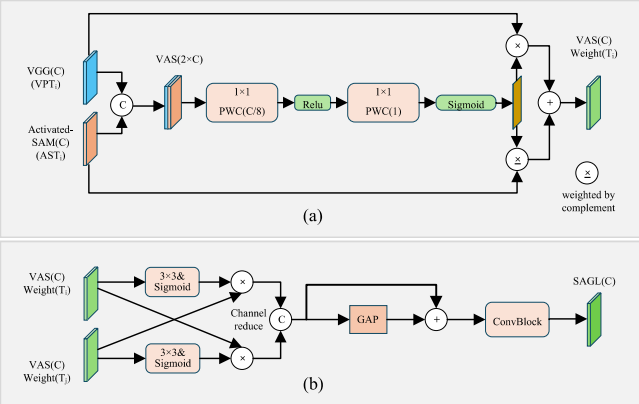

Fig. 4. The architectural framework of SAGL (DEFEAM(a), BIGLGM(b)).

C. SAM-Assisted Global-Local Fusion Module

After obtaining the multi-level features extracted by both the main encoder and the auxiliary encoder, they are fed into the Dual-Encoder Feature Enhancement Attention Module (DEFEAM) within the SAGL Fusion Module in a one-to-one correspondence. As illustrated in Fig. 4(a), this module first concatenates features from the same temporal phase and same dimension but different encoders. It then applies two 1×1 pointwise convolutions to enhance the "primary-auxiliary fused features", reducing the channel dimensions from 2×C $\rightarrow$ C/8 $\rightarrow$ 1, ultimately obtaining the dual-encoder feature enhanced attention weights $VASA \in [0,1]$. When $VASA$ approaches 1, features extracted by the VGG encoder are considered more important; when it approaches 0, features from the A-S encoder are prioritized. These features are then respectively weighted by the original $VASA$ value and its complement (1 - $VASA$), and summed to obtain the VAS Weight($VASW$). This process can be formally expressed as:

$$VASA_{Ti} = \sigma\left(1 \times 1\left(Relu\left(1 \times 1\left(Cat\left(VP_{Ti}, AS_{Ti}\right)\right)\right)\right)\right) \tag{6}$$

$$VASW_{Ti} = VP_{Ti} \otimes VASA_{Ti} + AS_{Ti} \otimes (1 - VASA_{Ti}) \tag{7}$$

where $\otimes$ denotes element-wise multiplication and σ represents the sigmoid function.

To enhance local-global information complementarity between bitemporal features, a Bidirectional Interaction Global-Local Gate Module (BIGLGM) is proposed, as illustrated in Fig. 4(b). After obtaining the $VASW$ for each temporal phase, bidirectional feature weighting and fusion are performed. Specifically, each $VASW$ is fed through a 3×3 convolution for weight extraction, activated by a sigmoid function, then multiplied with the $VASW$ from the other temporal phase. The resulting features are finally concatenated to form the Bidirectional Interaction Gate features BIG. The two BIG features are concatenated and then channel-wise reduced to match the original input size. Subsequently, Global Average Pooling is applied to capture global information, which is added to the original features (regarded as local information). This is followed by a ConvBlock containing two channel-unchanged 3×3 convolutional layers with ReLU activation to further enhance the feature output, ultimately producing the dual-encoder local-global fused features $SAGL$. The entire process can be described as follows:

$$\begin{aligned} BIG = Cat(&\sigma(Conv3 \times 3(VASW_{T_i}) \times VASW_{T_j}), \\ &\sigma(Conv3 \times 3(VASW_{T_j}) \times VASW_{T_i}) \end{aligned} \tag{8}$$

$$SAGL = ConvBlock(GAP(BIG) + BIG) \tag{9}$$

The final output is generated by progressively upsampling the deepest features, concatenating them layer by layer and decoded by a convolution Block (DConv). Taking the final concatenation process as an example, this procedure can be expressed as:

$$Output = DConv(Cat(Upsample(SAGL_2), SAGL_1)) \tag{10}$$

D. Loss

Since change detection is essentially a binary classification task, cross-entropy is employed as the loss function, which is defined as:

$$Loss = -\left(\hat{y}_i \log\left(y_i\right) + \left(1 - \hat{y}_i\right) \log\left(1 - y_i\right)\right) \tag{11}$$

where y_i and $\hat{y}_i$ represent the ground truth label and expected value (predicted probability) for pixel i, respectively.

III. Experiments

A. Datasets and Environment Setting

The LEVIR-CD [16] and WHU-CD [17] datasets are both publicly available building change detection datasets with image pairs uniformly cropped to 256×256 pixels. In LEVIR-CD, 890, 128, and 128 pairs are used for training, validation, and testing, respectively. For WHU-CD, 1,260 pairs are allocated for training and 690 for testing, with 10% of the training set randomly selected for validation. The SYSU-CD [18] dataset was employed to validate its generalization capability.

TABLE I
BRIEF INTRODUCTION OF ALL COMPARISON EXPERIMENTS

Year	Method	Brief Introduction
2021	BITNet	CNNs and transformers network captures temporal features as tokens, models context, and refines features with decoder.
2022	ICIF-Net	Intra-scale and inter-scale fusion network processes bi-temporal images with CNNs and transformers.
2023	HCGMNet	Change detection using a hierarchical change guiding map network.
2023	DMINet	Two-branch network uses attention to focus on real changes.
2024	SEIFNet	This architecture uses a Siamese network with hierarchical encoder and decoder integrating multiscale features.
2024	DGMA²-Net	This architecture uses Siamese CNN with MDFM fusion and DAM,DEAM refinement.
2025	ChangeTitans	A memory-augmented Transformer for efficient long-range modeling and pseudo-change suppression.
2025	SFEARNet	A SegFormer-based hybrid network using semantic flow and edge-aware refinement for detail optimization.

TABLE II
QUANTITATIVE COMPARISONS ON CD RESULTS. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Year		2021	2022	2023	2023	2024	2024	2025	2025	
Methods		BITNet	ICIFNet	HCGMNet	DMINet	SEIFNet	DGMA²-Net	ChangeTitans	SFEARNet	Ours
LEVIR (%)	F1	89.18	90.33	91.76	90.61	89.71	91.14	89.87	90.74	92.18
	Pre	90.89	91.15	93.10	92.56	92.23	91.93	90.89	91.94	93.11
	Rec	86.63	89.53	90.46	88.74	87.32	90.36	89.15	89.58	91.28
	IoU	80.48	82.36	84.78	82.83	81.61	83.72	81.34	83.06	85.50
WHU (%)	F1	84.42	88.97	92.08	89.92	87.30	93.09	92.55	93.26	94.84
	Pre	83.54	88.68	93.93	89.97	91.81	94.50	94.57	94.33	95.23
	Rec	85.31	89.26	90.31	89.87	83.21	91.89	90.62	92.05	94.44
	IoU	73.03	80.13	85.33	81.69	77.46	87.08	86.14	87.38	90.18

ASFCD is implemented in the Pytorch framework, and all experiments were tested on Intel Xeon Gold 5315Y(CPU)and NVIDIA A800(GPU). Using the AdamW optimizer, we reduced the loss by setting the weight decay and learning rate to 0.0025 and 0.0005, respectively. To facilitate easier comparison with other state-of-the-art methods, all data were cropped to 256×256. Additionally, data augmentation techniques were employed to enhance the data,including random Gaussian noise, random salt-and-pepper noise, random cropping, and random rotation at certain angles. Batch size and maximum epochs were selected as 8 and 50 for the experiments to allow the model to converge based on several numbers despite the limitations of the GPU. Over the total of 50 training epochs, we saved the model with the best F1 score and IoU on the validation set for testing.

Evaluation Metrics: Higher F1-scores and IoU generally indicate better change detection performance in experiments. Additionally, Precision and Recall are also important benchmark metrics. They can all be used to compare the predicted maps with the GT and are all explicitly defined as follows:

$$\text{F1} = \frac{2}{\text{Pre.}^{-1} + \text{Rec.}^{-1}} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{Precison} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{IoU} = \frac{TP}{TP + FN + FP} \tag{15}$$

the letters TP, TN, FP, and FN represent the respective sums of true positives, true negatives, false positives, and false negatives. Higher F1 and IoU values indicate better CD performance in the experiments.

B. Comparison With State-of-the-Art Methods

1) Introduction to Comparative Methods: To verify the effectiveness of the ASFCD, we selected several representative networks from recent years for comparative experiments. SEIF-Net [14] is a network based on convolution and attention mechanisms. DMINet [19] is a network based on convolution, attention mechanisms, and Transformer. BIT-Net [20], ICIF-Net [21] and SFEARNet [15] are networks based on convolution and Transformer. DGMA²-Net [22] and HCGMNet [23] are based on multi-scale, convolution, and attention mechanisms. ChangeTitans [24] Enhances standard Transformers with an external neural memory module for long-range context modeling. The specific introductions of the above eight methods are shown in Table I. To ensure the fairness of the comparison, they were replicated using their published source code and default hyperparameters, and the training epochs were standardized.

2) Analysis of Experimental Results: To validate the effectiveness of the ASFCD network, we selected several representative networks from recent years for comparative experiments on the LEVIR-CD and WHU-CD datasets, and the comparative results are summarized in Table II. ASFCD demonstrates significantly superior change detection performance on both datasets. On both the LEVIR-CD and WHU-CD datasets, ASFCD achieves the best results across all metrics. On LEVIR-CD, it outperforms the second-best model

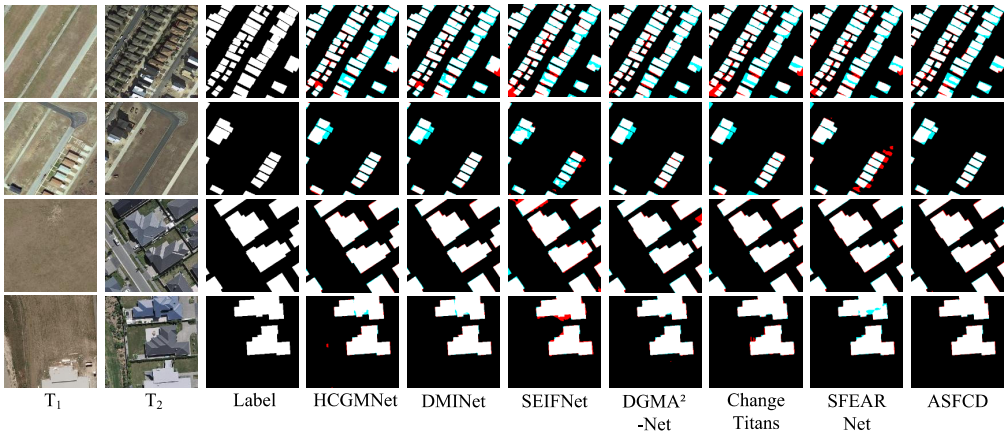

Fig. 5. Visualization of the results on the LEVIR-CD and WHU-CD datasets. TP, TN, FN, and FP are denoted as white, black, blue, and red.

by more than 0.7% in both Recall and IoU, while on WHU-CD, it exceeds the second-best model by 1.58% in F1-score and 2.8% in IoU. These advantages indicate that the dual-encoder architecture integrated with SAM can more effectively capture both local and global features compared to other encoder combination schemes. Furthermore, replacing MLPs with NewActNet enhances Vision Transformer's ability to model long-range dependencies and enables more flexible feature processing.

TABLE III
ABLATION STUDIES OF EACH MODULE, AF DENOTES THE ACTIVATION FUNCTION AND PARA MEANS PARAMETERS.

Ablation	LEVIR-CD(%)		WHU-CD(%)		para(M)
	F1	IoU	F1	IoU	
only VGG	91.11	83.67	91.25	83.90	21.32
only A-S	88.71	79.71	89.78	81.46	57.04
VGG+SAGL	91.69	84.65	93.18	87.22	30.85
A-S+SAGL	89.11	80.35	89.24	80.57	69.17
original SAM	91.81	84.86	94.20	89.04	151.54
NA for AF	91.97	85.13	94.33	89.28	94.36
ASFCD	92.18	85.50	94.84	90.18	84.04

Fig.5 presents a visual comparison between ASFCD and other networks. The first two rows display detection results in dense or small change areas, where ASFCD shows almost no false detections (red zones). This demonstrates its superior performance in detecting small and densely distributed changes compared to other methods such as SFEARNet, while also producing fewer missed detections (blue zones)—a crucial advantage particularly in dense change regions. The latter two rows involve detecting relatively sparse and large change areas. The change regions detected by ASFCD align almost perfectly with the ground truth, thus achieving exceptionally high F1-score and IoU. In contrast, other networks exhibit issues such as significant false detections and extensive missed detections, resulting in considerably inferior performance metrics. These results further demonstrate that ASFCD, equipped with a dual-encoder structure and local-global integration capability, achieves outstanding performance in change detection tasks.

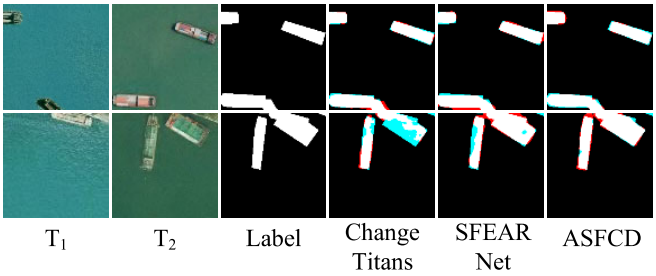

Fig. 6. Visualization of the results on the SYSU-CD datasets.

C. Ablation Study and Generalization Study

Our research focuses on three key aspects of ASFCD: the main encoder selection, NewActNet's replacement position in the Transformer, and dual-encoder fusion efficacy. Therefore, extensive ablation studies were conducted on ASFCD, and the experimental results are presented in Table III. As shown in the first two rows, VGG actually achieves better baseline performance than Activated-SAM due to its exceptional detail preservation capability. This is precisely why our method employs Activated-SAM as the auxiliary encoder to compensate for VGG's limitations in handling long-range interactions. Moreover, it can be observed that while using Activated-SAM alone for feature extraction yields suboptimal results, its combination with the VGG encoder through the SAGL fusion module leads to a notable improvement. Even employing the original SAM's ViT structure alone can achieve decent performance, albeit at the cost of a sharp increase in parameter count. Consequently, NewActNet was incorporated and the following experiments were conducted: First, the GELU activation function in the Transformer was replaced with NewActNet to serve as the actual activation function. As shown in Table III, this indeed reduced the parameter count and yielded some improvement in metrics, though the performance remained at a moderate level. Therefore, based on the inherent principle of ActNet (a novel structure capable of replacing both KANs and MLPs), we attempted to directly

replace the MLP layers. The results show that this approach not only reduced the parameter count to approximately half of the original (from 151.54M to 84.04M, representing a reduction of 44.54%) but also achieved significant performance gains on both datasets, particularly on WHU-CD where IoU increased by 1.14%, all without requiring significantly higher computational costs during actual operation. Additionally, the SAGL module effectively integrates features extracted by the dual encoders and successfully processes local-global information, leading to substantial performance improvements across both encoder-based architectures.

ASFCD was evaluated on the SYSU-CD dataset to assess its generalization capability. It achieved an F1-score of 82.63% and IoU of 70.40%, demonstrating robust performance across varied change scenarios. Fig.6 demonstrates that ASFCD produces fewer error zones compared to comparison networks.

IV. Conclusion

Inspired by the SAM, a remote sensing change detection network named ASFCD employs a dual encoder architecture with VGG and Activated-SAM. To address the main encoder VGG's limited long-range information processing capability, a novel Activated-SAM encoder based on the ViT component of the SAM encoder is proposed, which effectively eliminates the excessive parameter burden caused by MLP layers while significantly improving long-range feature interaction performance. To effectively fuse bitemporal and dual-encoder features while enhancing local-global information integration, the SAGL Fusion module is proposed. Extensive experiments were conducted on three public datasets to validate its detection accuracy and generalization capability. Future work will further investigate potential designs to replace MLP and explore whether NewActNet can fully substitute both MLP and KAN networks while delivering strong performance.

References

[1] Robert E Kennedy, Philip A Townsend, John E Gross, Warren B Cohen, Paul Bolstad, YQ Wang, and Phyllis Adams, "Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects," *Remote sensing of environment*, vol. 113, no. 7, pp. 1382–1396, 2009.

[2] Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sensing of Environment*, vol. 265, pp. 112636, 2021.

[3] Wei Wang, Cong Liu, Guanqun Liu, and Xin Wang, "Cf-gcn: Graph convolutional network for change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[4] Wei Wang, Luocheng Xia, and Xin Wang, "Building change detection based on fully convolutional network in high-resolution remote sensing images," in *International Conference on Intelligent Computing*. Springer, 2024, pp. 111–123.

[5] Wei Wang, Luocheng Xia, and Xin Wang, "Afpf-net: Adjacent-level feature progressive fusion full convolutional network for remote sensing change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[7] Zhiyuan Yan, Junxi Li, Xuexue Li, Ruixue Zhou, Wenkai Zhang, Yingchao Feng, Wenhui Diao, Kun Fu, and Xian Sun, "Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[8] Jie Zhang, Yunxin Li, Xubing Yang, Rui Jiang, and Li Zhang, "Rsam-seg: A sam-based model with prior knowledge integration for remote sensing image semantic segmentation," *Remote Sensing*, vol. 17, no. 4, pp. 590, 2025.

[9] Enkai Zhang, Jingjing Liu, Anda Cao, Zhen Sun, Haofei Zhang, Huiqiong Wang, Li Sun, and Mingli Song, "Rs-sam: integrating multi-scale information for enhanced remote sensing image segmentation," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 994–1010.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[12] Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, and Santiago Pourteau, "Convolutional kolmogorov-arnold networks," *arXiv preprint arXiv:2406.13155*, 2024.

[13] Leonardo Ferreira Guilhoto and Paris Perdikaris, "Deep learning alternatives of the kolmogorov superposition theorem," *arXiv preprint arXiv:2410.01990*, 2024.

[14] Yanyuan Huang, Xinghua Li, Zhengshun Du, and Huanfeng Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[15] Miao Li, Dongping Ming, Lu Xu, Dehui Dong, and Yu Zhang, "Sfear-net: A network combining semantic flow and edge-aware refinement for highly efficient remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

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

[17] Shunping Ji, Shiqing Wei, and Meng Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on geoscience and remote sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[18] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021.

[19] Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[20] Hao Chen, Zipeng Qi, and Zhenwei Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[21] Yuchao Feng, Honghui Xu, Jiawei Jiang, Hao Liu, and Jianwei Zheng, "Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[22] Zilu Ying, Zijun Tan, Yikui Zhai, Xudong Jia, Wenba Li, Junying Zeng, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti, "Dgma 2-net: a difference-guided multiscale aggregation attention network for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[23] Chengxi Han, Chen Wu, and Bo Du, "Hcgmnet: A hierarchical change guiding map network for change detection," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 5511–5514.

[24] Zhenyu Yang, Gensheng Pei, Yazhou Yao, Tianfei Zhou, Lizhong Ding, and Fumin Shen, "Changetitans: Towards remote sensing change detection with neural memory," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.