# Activated Three Stages Building Change Detection Network

Xin Wang[1], Aocheng Shu[2], and Wei Wang[1(✉)]

[1]School of Physics and Electronic Science, Changsha University of Science and Technology,
[2]School of Computer Science and Technology, Changsha University of Science and Technology,
Changsha 410114, China
wangxin@csust.edu.cn

**Abstract.** Current Building Change Detection methods suffer from several drawbacks, including inadequate multi-scale feature interaction, blurred boundaries, and excessive computational costs. To address these challenges, ActPML is proposed. It introduces three hierarchically coupled stages: 1) a Pre-ActNet stage with channel-wise fusion and a novel Global Activation Layer that dynamically filters features while avoiding KAN's heavy computations; 2) a Middle stage composed of Global-Local Feature Aggregator and Self Deep Fusion Enhancer for bi-temporal global-local deep refinement; 3) a Late stage with Global-Local Self Enhancer to guide the decoder. ActPML achieves state-of-the-art performance on the LEVIR-CD and WHU-CD datasets, with F1-scores reaching 92.14% and 94.43%. Additionally, generalization capability was further validated on the CLCD dataset. Moreover, the Global Activation Layer significantly reduces computational overhead compared to KANs and MLPs, requiring substantially less runtime and fewer parameters. Our code can be seen at https://github.com/zcbg123/ActPML.git.

**Keywords:** BCD, stage, global-local, KAN, GAL.

## 1 Introduction

Traditional Building Change Detection (BCD) methods may introduce substantial noise and pseudo-changes in the resulting change maps. With breakthroughs in deep learning, Convolutional Neural Networks (CNNs) have significantly improved feature extraction and classification accuracy, while substantially reducing the reliance on manual intervention and lowering the overall workload compared to traditional methods. Consequently, numerous fully convolutional remote sensing change detection networks based on CNNs have emerged [1, 2, 3]. In CNN-based change detection networks, features from different levels undergo interaction and integration after extraction, with the most prevalent approaches being previous-stage fusion [4] and late-stage fusion [5]. In previous-stage fusion, bi-temporal images are typically processed through simple concatenation or differencing after feature extraction. Despite achieving reasonably satisfactory results, these methods lack refined modeling of cross-temporal feature interaction.

As shown in Fig.1, useful information is merely subjected to simple concatenation or differencing operations after extraction [6], consequently introducing considerable noise. Additionally, although multi-scale feature fusion mechanisms [7, 8] have been widely adopted in deep learning, inadequate interaction across hierarchical features persists, ultimately leading to suboptimal performance.

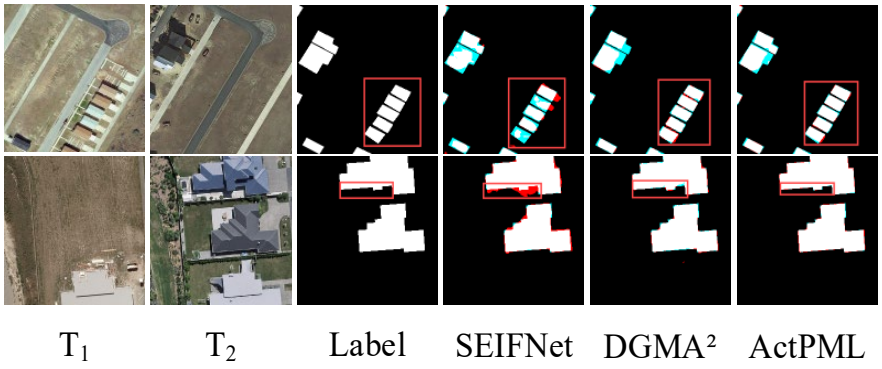


**Fig. 1.** Visualization results of different change detection methods. TP, TN, FN, and FP are represented by white, black, blue, and red, respectively.

Furthermore, the introduction of Kolmogorov-Arnold Networks(KANs) [9, 10] has pioneered a new research direction in deep learning. KANs provides a novel solution compared to multi-layer perceptrons(MLPs) [11] that struggle with long-sequence processing. Although KANs address the issue of excessive parameter numbers, they impose stringent hardware requirements and incur substantially higher training time costs. Several researchers are actively exploring solutions to these challenges, such as ActNet [12]. Although ActNet effectively mitigates the training cost issue, its potential for broad applicability across deep learning remains uncertain.

To address these issues, a multi-scale fusion network named ActPML is proposed, which leverages features from the three stages. The network employs a global activation layer as the activation mechanism for previous-stage features, enabling dynamic feature filtering. The processed features are subsequently forwarded to the middle and late stages, where Global-Local Feature Aggregator, Self Deep Fusion Enhancer and Global-Local Self Enhancer modules perform local-global and self-deep multi-scale feature enhancement and refinement, ultimately producing the change map. The main contributions of our work are as follows:

(1)To address the insufficient cross-stage feature interaction caused by simple concatenation or differencing operations in existing methods, a hierarchical three-stage architecture is proposed. In this architecture, learnable cross-scale channel adaptation and dynamic feature filtering are performed in the Pre-ActNet stage, bi-temporal global-local deep fusion is realized by coupling the Global-Local Feature Aggregator (GLFA) and the Self Deep Fusion Enhancer (SDFE) in the Middle stage, and self-guided refinement is achieved through the Global-Local Self Enhancer (GLSE) in the Late stage.

(2)To address the accuracy-efficiency trade-off of activation functions, where KANs incur prohibitive computational costs and MLPs suffer from fixed-weight filtering, a Global Activation Layer (GAL) based on NewActNet is developed. This module is validated for the first time under the theoretical guarantee of Laczkovich's theorem, and is shown to achieve comparable or superior accuracy to KANs while being significantly faster in inference and requiring fewer parameters.
(3)To provide comprehensive experimental validation that is lacking in prior work, extensive ablation studies are conducted on GAL placement, three-stage contributions, backbone choices, and the individual effects of GLFA and SDFE. The proposed ActPML is compared against eight state-of-the-art methods on LEVIR-CD and WHU-CD, and its strong generalization capability is quantitatively demonstrated on the CLCD dataset with F1 and IoU metrics.

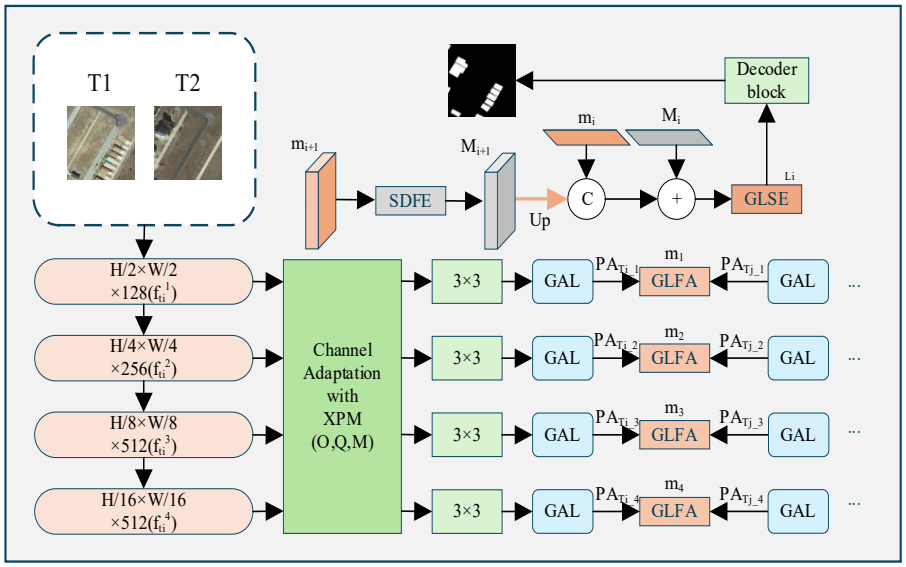


**Fig. 2.** The architectural framework of ActPML.

## 2 Method

### 2.1 Overall Structure

Fig.2 illustrates the overall network architecture of ActPML. The bi-temporal images with dimensions H×W×3 are fed into a VGG-16 backbone network for multi-scale feature extraction, where the images are progressively downsampled by a factor of two to H/16×W/16 resolution, with channel dimensions successively increasing to 128, 256, 512, and 512. In the previous stage, comprising Channel Adaptation and 3×3 convolu-

tional layers, integrates single-temporal features from different scales. This allows features at each scale to gain rich semantic information supplementation from other scales, which are then processed through the GAL module containing NewActNet to yield P-A features. To capture both global-local feature interactions and deep self-information integration, the P-A features are divided into bi-temporal single-scale feature groups and processed through the middle stage containing GLFA and SDFE. This yields mid-stage-pre features and mid-stage-late features. During the late-stage phase, the mid-stage-pre features guide the mid-stage-late features through GLSE, ultimately producing the change map.

### 2.2 Pre-ActNet stage Block

To enable feature interaction and fusion across different scales, the Pre-ActNet stage Block is designed incorporating Channel Adaptation, convolutional layers, and a GAL module, as illustrated in Fig. 2. The Channel Adaptation comprises four channel transformation branches with similar functions, as exemplified by the branch functionality shown in Fig. 3(a). Each branch integrates and enhances features at its assigned scale, enabling cross-scale fusion with information from other scales. Taking the fourth branch as an example, let $f_{ti}^{j}$ (where i ∈{1,2}, j ∈{1,2,3,4}) denote the multi-scale features extracted from the backbone network. Given that $f_{ti}^{4}$ has a channel dimension of 512, $f_{ti}^{1}$ is first processed by an Octuple Pooling Module with channel adjustment, followed by a 3×3 convolutional layer to align its spatial resolution with $f_{ti}^{4}$. Similarly, $f_{ti}^{2}$ passes through a Quadruple Pooling Module with channel adjustment and a 3×3 convolutional layer, while $f_{ti}^{3}$ is adjusted via a Max Pooling Module and a 3×3 convolutional layer. For $f_{ti}^{4}$, which requires fusion with other scales, it is processed by separate 3×3 and 1×1 convolutional layers for channel adaptation. Subsequently, concatenation and addition operations are performed to generate $p_{ti\_4}$. Subsequently, all features that have passed through the 3×3 convolutional layers are concatenated to fuse information from all scales, and then refined by a final 3×3 convolutional layer to obtain feature $P_{ti\_4}$. This operation enables deep-level features to incorporate high-resolution information from shallow-level features. The complete feature transformation process can be summarized as:

$$p_{ti\_4} = Conv1 \times 1(f_{ti}^{4}) + Cat(XPM(f_{ti}^{1}, f_{ti}^{2}, f_{ti}^{3}, f_{ti}^{4})) \quad (1)$$

$$P_{ti\_4} = Conv3 \times 3(p_{ti\_4}) \quad (2)$$

where The XPM (cross-scale pooling module) conditionally applies one of the following operations depending on the input scale: Octuple Pooling (three cascaded max-pooling, stride 2 each), Quadruple Pooling (two cascaded max-pooling), Max Pooling (single pooling), or Identity (no pooling).

Following the acquisition of pre-stage features, the Global Activation Layer (GAL) is designed to dynamically filter these features. The GAL (Fig.3(c)) dynamically filters pre-stage features via NewActNet, comprising three NewActLayer and LayerNorm groups. Features are activated by NewActLayer, which operates akin to MLPs with

learnable activation functions $\phi_i$ per head, followed by linear layers $\lambda_i$ (output dim=1). All heads'outputs are concatenated for downstream processing. NewActLayer serves as a lightweight alternative to KANs with equivalent functionality. The design of NewActLayer is grounded in Theorem 1.2 established by Laczkovich [13], which states: Let $C(\mathbb{R}^d)$ denote the space of continuous functions on $\mathbb{R}^d$, where d and $m > (2+\sqrt{2})(2d-1)$ are integers. There exist positive constants $\lambda_{ij} > 0$ (j=1,...,d; i=1,...,m) and m continuous monotonically increasing functions $\phi_i \in C(\mathbb{R})$(i=1,…,m) such that for every bounded function $f \in C(\mathbb{R}^d)$, there exists a continuous function $g \in C(\mathbb{R})$ satisfying:

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} g(\boldsymbol{\lambda}_i \cdot \phi_i(\boldsymbol{x})) \tag{3}$$

where $\phi_i$ are applied element-wise and $\lambda_i = (\lambda_{i1},\ldots,\lambda_{id}) \in \mathbb{R}^d$.

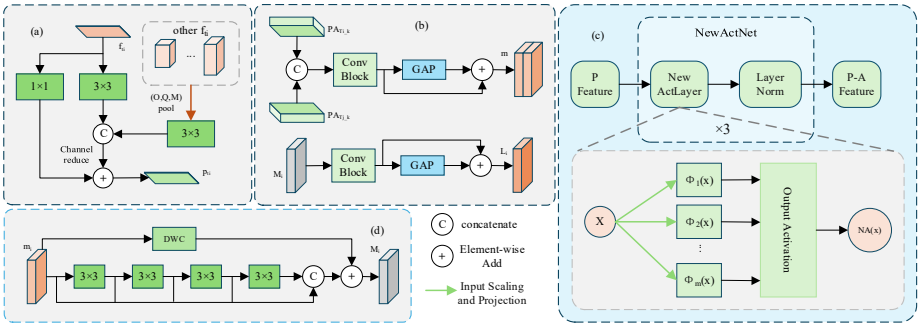


**Fig. 3.** The architectural framework of Channel Adaptation(a), GLFA and GLSE(b), GAL(c), SDFE(d).

Expressing the entire process in matrix form, let $\boldsymbol{\Phi}_{ij} = \phi_i(x_j)$ and $\boldsymbol{\Lambda}_{ij} = \lambda_{ij}$, and define S: $\mathbb{R}^{a\times b} \to \mathbb{R}^a$ as the function returning row sums of a matrix. Building upon this framework, we redesign the operational workflow of NewActLayer, which now requires only O(d) internal functions instead of $O(d^2)$ in KANs, thereby significantly reducing computational overhead. The complete operation of NewActLayer can be formulated as:

$$NewActLayer_{\boldsymbol{\Lambda}}(\boldsymbol{x}) = S(\boldsymbol{\Lambda} \odot \boldsymbol{\Phi}(x)) \tag{4}$$

$$NA(\mathrm{x}) = Normalization(NewActLayer(\mathrm{x})) \tag{5}$$

$$GAL(\mathrm{x}) = NA(NA(NA(\mathrm{X}))) \tag{6}$$

where $\odot$denotes the Hadamard product, and NA refers to the module group composed of the NewActLayer and LayerNorm layers.

Therefore, when pre-stage features are input into the GAL, GAL dynamically filters and activates the features in a manner analogous to MLPs layers or KANs. The process can be formulated as follows:

$$PA_{Ti} = GAL(P_{Ti}) \tag{7}$$

### 2.3 Mid-stage Block

Since the features extracted by Pre-ActNet stage Block only involve fusion and interaction of multi-scale features within individual temporal phases, establishing bi-temporal interaction becomes particularly important for enhancing information extraction in later stages. Therefore, the Mid-stage Block is designed (Fig.3(b), (d))—comprising GLFA and SDFE—to process dual-temporal single-scale features through a convolution block (including 3×3 convolution, ReLU, and Batch Normalization). The GLFA utilizes Global Average Pooling for global vectors and a convolution-normalization-activation chain for local contexts. Reweighted features are summed with original P-A features and bi-temporally concatenated, generating mid-stage-pre features $m_i$ that adaptively balance cross-temporal interactions and mitigate imaging interference, as illustrated in Fig.3(b). For inputs $T_1$ and $T_2$, the output can be expressed as:

$$\mathrm{m}_i = GLFA(PA_{T1\_i}, PA_{T2\_i}) \tag{8}$$

$$GLFA(\mathrm{X}_1, \mathrm{X}_2) = Global\ (Cat(\mathrm{X}_1, \mathrm{X}_2)) + Local\ (Cat(\mathrm{X}_1, \mathrm{X}_2)) \tag{9}$$

$$Global(\mathrm{x}) = Relu(BN(Conv3 \times 3\ (Gap(x)))) \tag{10}$$

$$Local(\mathrm{x}) = Relu(BN(Conv3 \times 3\ (x))) \tag{11}$$

where Gap denotes Global Average Pooling, and BN refers to Batch Normalization.

To enhance temporal change capture, SDFE is proposed, employing four parallel convolutional paths for hierarchical multi-scale extraction. Each path processes base features from preceding stages, supplements hierarchical features via concatenation, and integrates them with the original input through a Depth-wise Convolution layer (DWC), followed by summation to generate mid-stage-late features $M_i$ The output can be formulated as:

$$M_i = SDFE(\mathrm{m}_i) \tag{12}$$

$$SDFE(x) = Cat(Conv3 \times 3(\mathrm{x}), Conv3 \times 3(Conv3 \times 3(\mathrm{x})), \dots) + DWC(\mathrm{x}) \tag{13}$$

where Cat internally represents the feature concatenation after multiple convolutions.

### 2.4 Late-stage Block

Unlike conventional methods relying on final-stage feature sum or concatenation, the representation is enhanced by progressive aggregation of mid-stage outputs $M_i$ from shallow to deep layers, followed by 1×1 convolution for change map generation. The Late-stage Block uniquely integrates Mid-stage Block-derived $m_i$ and $M_i$ features for mutual complementarity: deeper $M_{i+1}$ features are upsampled, concatenated with preceding $m_i$, interact with prior $M_i$ via summation, and The features are processed

through GLSE and subsequently decoded into $L_i$ (Fig.2) via decoder block (3×3 and 1×1 convolutions).

Every $L_i$ undergoes self-enhanced refinement via the GLSE (Fig.3(b)), providing auxiliary guidance. GLSE shares functional similarity with GLFA but differs in input scope: GLFA handles bi-temporal data, while GLSE operates on single-temporal inputs. The complete workflow can be formulated as follows:

$$L_i = GLSE(Cat(Upsample(M_{i+1}),m_i)+M_i) \tag{14}$$

$$Out = Conv1 \times 1(Conv3 \times 3(L_i)) \tag{15}$$

### 2.5 Loss

Building change detection constitutes a binary classification task. To mitigate class imbalance and boundary ambiguity inherent in this task, the cross-entropy loss function is employed to enhance prediction accuracy. The cross-entropy loss is defined as:

$$Loss = -(\tilde{y}_i \log(y_i) + (1 - \tilde{y}_i) \log(1 - y_i)) \tag{16}$$

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

**Table 1.** Quantitative comparisons on LEVIR-CD and WHU-CD results. The best results are highlighted in bold, respectively.

| Model | LEVIR-CD | WHU-CD |
|---|---|---|
| | F1./Pre./Rec./IoU.(%) | F1./Pre./Rec./IoU.(%) |
| BITNet(2021) | 89.18/91.89/86.63/80.48 | 84.42/83.54/85.31/73.03 |
| ChangeFormer(2022) | 90.17/91.36/89.02/82.10 | 91.62/94.73/88.71/84.54 |
| DMINet(2023) | 90.61/92.56/88.74/82.83 | 89.92/89.97/89.87/81.69 |
| EATDer(2024) | 91.20/91.74/90.67/83.82 | 90.01/91.32/88.74/81.83 |
| SEIFNet(2024) | 89.71/92.23/87.32/81.34 | 87.30/91.81/83.21/77.46 |
| DGMA$^2$-Net(2024) | 91.14/91.93/90.36/83.72 | 93.09/94.50/91.89/87.08 |
| SFEARNet | 90.74/91.94/89.58/83.06 | 93.26/94.33/92.05/87.38 |
| ChangeTitans | 89.87/90.89/89.15/81.34 | 92.55/94.57/90.62/86.14 |
| ActPML(ours) | **92.14/93.16/91.15/85.43** | **94.43/95.65/93.24/89.45** |

## 3 Experiments

### 3.1 Datasets and Environment Setting

The LEVIR-CD [14] and WHU-CD [15] datasets are publicly available for building change detection. Both datasets use 256×256-pixel patches: LEVIR-CD is divided into

890, 128, 128 for train, val, test, while WHU-CD employs 1260 and 690 for train and test with 10% of training data as validation. ActPML is implemented in PyTorch, trained and tested on an NVIDIA A800 GPU using AdamW optimizer (weight decay: 0.0025, learning rate: 0.0005). Data augmentation includes Gaussian and salt-and-pepper noise, random cropping and rotation. Training employs batch size 8 over 50 epochs, balancing resource efficiency and performance. The model achieving highest validation F1 or IoU is retained for evaluation.

Evaluation metrics include F1-score, Precision(Pre.), Recall(Rec.), and IoU, where higher F1 or IoU values indicate superior detection performance. They are explicitly defined as follows:

$$F1 = \frac{2}{Pre.^{-1} + Rec.^{-1}} \tag{17}$$

$$Rec. = \frac{TP}{TP+FN} \tag{18}$$

$$Pre. = \frac{TP}{TP+FP} \tag{19}$$

$$IoU = \frac{TP}{TP+FN+FP} \tag{20}$$

the letters TP, TN, FP, and FN represent the respective sums of true positives, true negatives, false positives, and false negatives.

### 3.2 Comparison With State-of-the-Art Methods

To evaluate ActPML, we compared it with key recent networks: SEIF-Net [16] (convolution and attention), DMINet [17] (convolution, attention, and Transformer), BIT-Net [18] and SFEARNet [19] (Transformer or hybrid Transformer-CNN), EATDer [8] (lightweight edge-aware adaptive Transformer), ChangeFormer [20] (multi-scale MLPs and Transformer), DGMA²-Net [21] (multi-scale convolution and attention), and ChangeTitans [22] (Transformer with external memory for long-range context).

Table 1 demonstrates the superiority of ActPML over SOTA networks (e.g., EATDer [8] and SFEARNet [19]), as it achieves the highest scores in F1, Recall, Precision, and IoU on both the LEVIR-CD and WHU-CD datasets. In particular, it outperforms the second-best methods by 1.61% and 2.07% in IoU on LEVIR-CD and WHU-CD, respectively. Existing approaches—including attention-based networks (such as DGMA²-Net [21]), and Transformer or hybrid Transformer-CNN models (such as ChangeFormer [20] and BITNet [18])—fail to effectively integrate dynamic filtering mechanisms or prioritize change-aware feature differentiation. In contrast, ActPML adaptively processes features through three hierarchical blocks: a single-temporal multi-scale fusion module with dynamic feature filtering, a bi-temporal single-scale fusion module, and an interactive mid-stage fusion module. This design enables an effective balance between context preservation and feature refinement.

Fig.4 presents visualizations of ActPML on the LEVIR-CD (rows 1, 2) and WHU-CD (rows 3, 4) datasets. Compared to competing methods, it significantly reduces both false positive (red) and false negative (blue) errors. On LEVIR-CD, ActPML detects

subtle changes in sparse regions (row 1) and reduces false alarms in dense building clusters (row 2), while effectively overcoming interference from vegetation and other non-building elements. On WHU-CD, it achieves near-complete detection in both dense and sparse areas, suppresses false positive artifacts (row 3), mitigates interference from vehicles (row 4) and other confusers, with minimal misdetection. These results validate that ActPML delivers precise change detection across both dense and sparse scenarios, demonstrates strong resilience to environmental interference, and achieves superior accuracy compared to existing methods.

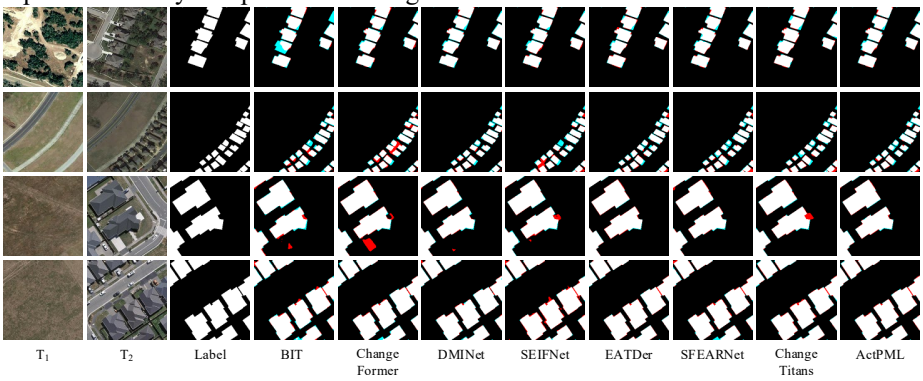


**Fig. 4.** Visualization of the results on the LEVIR-CD and WHU-CD datasets. TP, TN, FN, and FP are denoted as white, black, blue, and red.

**Table 2.** Ablation study on GAL(Times denotes the testing time during model evaluation and para means parameters.)

| Model | LEVIR-CD | WHU-CD | Parameters(M) |
|---|---|---|---|
| | F1.(%)/Times(s) | F1.(%)/Times(s) | |
| P-MLP | 74.36/43.1 | 74.50/**27.6** | 156.16 |
| P-A AL | 91.94/51.3 | 93.84/50.2 | **54.04** |
| P-A ALAL | 91.98/52.0 | 93.91/51.3 | 54.12 |
| M-A ALALAL | 91.92/111.4 | 94.22/49.5 | 54.91 |
| L-A ALALAL | 91.88/78.1 | 94.14/35.5 | 54.92 |
| P-A ALALALAL | 92.02/124.7 | 94.23/48.9 | 55.01 |
| P-KANConv | **92.15**/1019.0 | 93.39/386.0 | 58.06 |
| ActPML | 92.14/**38.7** | **94.43**/33.9 | 54.20 |

### 3.3 Ablation Experiments

Ablation of GAL: To explore the application of GAL in change detection, we integrated GAL at different hierarchical stages and compared it with MLPs or KAN convolutional networks by adjusting the configuration of ActLayer-LayerNorm (AL). As shown in Table 2, due to the use of fixed-weight filtering in the dynamic preprocessing stage, the

MLPs model exhibits low performance. KAN convolution slightly outperforms ActPML on LEVIR-CD but shows unstable performance on WHU-CD with higher inference costs. Moreover, GAL employs fewer parameters (54.2M) than both MLPs (156.16M) and KAN convolutions (58.06M). Furthermore, the test time is significantly lower than that of KAN convolution and is almost comparable to that of the MLP model. Therefore, GAL surpasses both in balancing accuracy and efficiency.

Further analysis of the placement order and number of AL modules reveals that adding LayerNorm after NewActLayer helps improve feature standardization. Once the number of AL groups exceeds three, performance actually declines—excessive filtering discards subtle features. Using three groups of AL in GAL achieves the optimal balance between selective filtering and feature preservation, striking a computational-performance equilibrium. Moreover, compared to the Mid-stage Block and Late-stage Block, GAL is more suitable in the Pre-ActNet Block. This is because the former two both perform activation operations on features, and repeated activation filtering can lead to excessive sparsification of features, thereby reducing detection effectiveness.

**Table 3.** Ablation Study on three stages.

| Model | LEVIR-CD | WHU-CD |
|---|---|---|
| | F1.(%)/IoU.(%) | F1.(%)/IoU.(%) |
| Baseline | 89.92/81.65 | 91.25/83.90 |
| Previous × Mid √ Late √ | 91.90/85.01 | 93.87/88.45 |
| Previous √ Mid × Late √ | 91.92/85.05 | 94.00/88.69 |
| Previous √ Mid √ Late × | 91.91/85.03 | 92.99/86.91 |
| Previous √ Mid × Late × | 91.78/84.81 | 93.93/88.55 |
| Previous × Mid √ Late × | 91.31/84.02 | 93.41/87.63 |
| Previous × Mid × Late √ | 91.79/84.83 | 93.80/88.32 |
| ActPML | **92.14/85.43** | **94.43/89.45** |

Ablation of three stages: To validate the individual contributions and interconnections of the previous, middle, and late stage blocks (P-B, M-B, L-B), ablation studies were conducted on three stages components, as summarized in Table 3. Removing the P-B resulted in significant drops in both F1 and IoU metrics, attributable to the loss of multi-scale fusion capability through CA. Notably, an isolated P-B failed to achieve effective bi-temporal interaction. The M-B, incorporating GLFA and SDFE, enables dual bi-temporal interaction and self-deep fusion. Its removal led to moderate performance decline, while using it alone severely degraded results, confirming its crucial role as a bridge between P-B and L-B. Eliminating the L-B (containing GLSE) had minimal impact on LEVIR-CD but substantially affected WHU-CD, demonstrating GLSE's capacity for guiding deep feature refinement. The suboptimal performance of an isolated L-B underscores its dependence on both the bi-temporal local-global deep interactions derived from GLFA and SDFE in M-B and the multi-scale fused features generated by CA in P-B.

**Table 4.** Ablation Study on GLFA and SDFE.

| Model | LEVIR-CD | WHU-CD |
|---|---|---|
| | F1.(%)/IoU.(%) | F1.(%)/IoU.(%) |
| None | 91.92/85.05 | 94.00/88.69 |
| GLFA√ SDFE× | 92.01/85.20 | 94.10/88.86 |
| GLFA× SDFE√ | 92.07/85.30 | 94.04/88.75 |
| ActPML | **92.14/85.43** | **94.43/89.45** |

Ablation of GLFA and SDFE. As shown in Table 4, on LEVIR-CD, removing both modules (None) gave 91.92% and 85.05%. Using only GLFA or only SDFE raised F1 to 92.01% and 92.07%, respectively, while their combination (ActPML) achieved the best result of 92.14%/85.43%. On WHU-CD, the baseline was improved to 94.10% and 88.86% with GLFA alone and to 94.04% and 88.75% with SDFE alone; the full model reached 94.43% and 89.45%. These results confirm that GLFA and SDFE are complementary, with their synergy being particularly beneficial for complex scenes like WHU-CD, where global-local aggregation is shown to be more influential than self-deep enhancement.

**Table 5.** Ablation Study on different backbones.

| backbone | LEVIR-CD | WHU-CD | parameters(M) |
|---|---|---|---|
| | F1./IoU.(%) | F1./IoU.(%) | |
| Resnet-18 | 90.97/83.44 | 89.33/80.72 | 66.33 |
| Resnet-34 | 90.78/83.12 | 90.49/82.62 | 92.58 |
| VGG-16 | **92.14/85.43** | **94.43/89.45** | **54.20** |

Ablation of Backbones: We conducted ablation studies on backbone networks by replacing VGG-16 with ResNet-18 and ResNet-34. As shown in Table 5, VGG-16 achieves higher accuracy with fewer parameters compared to these alternatives. For the ActPML network, which prioritizes feature extraction and fusion, VGG-16 more effectively captures image details than the ResNet variants. This outcome was achieved without relying on deeper architectures for training, thus yielding more pronounced improvements. Consequently, VGG-16 remains the optimal backbone network for ActPML.

### 3.4 Generalization Study

Fig.5 visualizes ActPML's performance on the CLCD [23] dataset. Although CLCD is not a building change detection dataset, ActPML still achieves nearly complete detection of change regions, demonstrating its strong generalization capability. As shown in Table 6, ActPML achieved an F1-score of 68.47% and IoU of 52.05%, demonstrating robust performance across varied change scenarios.

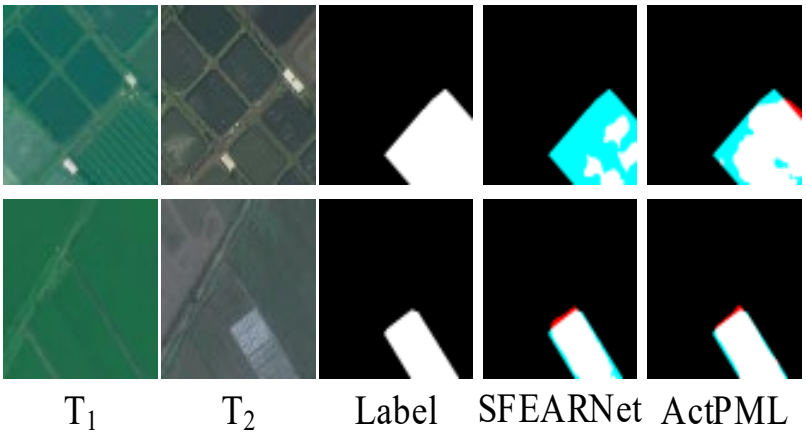


**Fig. 5.** Visualization of the results on the CLCD dataset.

**Table 6.** Generalization Study on CLCD dataset.

| Model | CLCD |
|---|---|
| | F1./IoU.(%) |
| SFEARNet | 67.85/51.35 |
| ActPML | **68.47/52.05** |

## 4 Conclusion

This study proposes ActPML, a novel three stages network, and explores whether ActNet—a potential alternative to MLPs or KANs—is suitable for change detection. It integrates three core stage blocks (P-A, M, and L), which employ Channel Adaptation, GLFA, SDFE, and GLSE to achieve bi-temporal multi-scale local-global deep feature extraction. To overcome the limitations of fixed parameters in MLPs and the high computational cost of KANs, a GAL module is designed, whose accuracy and computational efficiency have been validated on two BCD datasets. Also, generalization capability was further validated on the CLCD dataset. In the future, our research will focus on developing more efficient architectures and further exploring optimized structures capable of replacing MLPs.

**Acknowledgments.** This study was funded by National Science Innovation Special Zone Project (2019XXX00701), National Key Basic Research Program Project(624XXXX0206), Key research and development projects of Hunan Province (2020SK2134), Natural Science Foundation of Hunan Province (2019JJ80105, 2022JJ30625).

**Disclosure of Interests**. The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Wang, W., Liu, C., Liu, G., Wang, X.: Cf-gcn: Graph convolutional network for change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024)
2. Wang, W., Xia, L., Wang, X.: Building change detection based on fully convolutional network in high-resolution remote sensing images. In: International Conference on Intelligent Computing. pp. 111–123. Springer (2024)
3. Wang, W., Xia, L., Wang, X.: 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 **17**, 13853–13865 (2024)
4. Li, Z., Tang, C., Liu, X., Zhang, W., Dou, J., Wang, L., Zomaya, A.Y.: Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–12 (2023)
5. Zhou, C., Zhang, H., Guo, H., Zou, Z., Shi, Z.: A late-stage bitemporal feature fusion network for semantic change detection. IEEE Geoscience and Remote Sensing Letters (2024)
6. Fang, S., Li, K., Shao, J., Li, Z.: Snunet-cd: A densely connected siamese network for change detection of vhr images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021)
7. Han, C., Wu, C., Du, B.: Hcgmnet: A hierarchical change guiding map network for change detection. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. pp. 5511–5514. IEEE (2023)
8. Ma, J., Duan, J., Tang, X., Zhang, X., Jiao, L.: Eatder: Edge-assisted adaptive transformer detector for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2023)
9. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024)
10. Bodner, A.D., Tepsich, A.S., Spolski, J.N., Pourteau, S.: Convolutional kolmogorov-arnold networks. arXiv preprint arXiv:2406.13155 (2024)
11. Zhang, C., Wang, L., Cheng, S., Li, Y.: Swinsunet: Pure transformer network for remote sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022)
12. Guilhoto, L.F., Perdikaris, P.: Deep learning alternatives of the kolmogorov superposition theorem. arXiv preprint arXiv:2410.01990 (2024)
13. Laczkovich, M.: A superposition theorem of kolmogorov type for bounded continuous functions. Journal of Approximation Theory **269**, 105609 (2021)
14. Chen, H., Shi, Z.: A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote sensing **12**(10), 1662 (2020)
15. Ji, S., Wei, S., Lu, M.: Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on geoscience and remote sensing **57**(1), 574–586 (2018)
16. Huang, Y., Li, X., Du, Z., Shen, H.: Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
17. Feng, Y., Jiang, J., Xu, H., Zheng, J.: Change detection on remote sensing images using dual-branch multilevel intertemporal network. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)
18. Chen, H., Qi, Z., Shi, Z.: Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2021)

19. Li, M., Ming, D., Xu, L., Dong, D., Zhang, Y.: Sfearnet: 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)
20. Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. pp. 207–210. IEEE (2022)
21. Ying, Z., Tan, Z., Zhai, Y., Jia, X., Li, W., Zeng, J., Genovese, A., Piuri, V., Scotti, F.: Dgma 2-net: a difference-guided multiscale aggregation attention network for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
22. Yang, Z., Pei, G., Yao, Y., Zhou, T., Ding, L., Shen, F.: Changetitans: Towards remote sensing change detection with neural memory. IEEE Transactions on Geoscience and Remote Sensing (2025)
23. Liu, M., Chai, Z., Deng, H., Liu, R.: A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 4297–4306 (2022)