Beyond Attention: Convolutional Global Context for Remote Sensing Change Detection

Zhenyu Yang[1,3], Gensheng Pei[2(✉)], Junzhu Mao[1,3], Xinhao Cai[1,3], Tao Chen[1,3], and Yazhou Yao[1,3(✉)]

[1] Nanjing University of Science and Technology, Nanjing, China
[2] Department of Electrical and Computer Engineering Sungkyunkwan University, Suwon, Korea
[3] State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, Nanjing, China
✉Corresponding authors: peigsh@skku.edu, yazhou.yao@njust.edu.cn
Code: https://github.com/NUST-Machine-Intelligence-Laboratory/ChangeGCC

Abstract. Remote sensing change detection is caught in a long-standing trade-off: reliable suppression of pseudo changes requires global contextual reasoning, while attention-based models incur prohibitive quadratic complexity, hindering deployment on resource-constrained UAV and satellite platforms. This paper proposes **ChangeGCC**, a fully convolutional, linear-time framework that delivers efficient global context for change detection. ChangeGCC core features two key components: **G**lobally **C**onditioned **C**onvolution that enable convolutional global routing, and a dual-temporal fusion scheme that enhances semantic consistency across scales while highlighting genuine structural changes. Extensive experiments on multiple benchmarks show that ChangeGCC achieves state-of-the-art performance, reaching **85.28% IoU** and **92.06% F1** on the LEVIR-CD benchmark, while substantially reducing parameters, FLOPs, and inference latency compared to attention-based counterparts.

Keywords: Change detection · Remote sensing · Fully convolution

1 Introduction

Remote Sensing Change Detection (RSCD) aims to identify pixel-wise changes between multi-temporal images, enabling fine-grained monitoring of land-cover evolution over time. In practice, RSCD models can be deployed on UAV and satellite platforms to rapidly compare historical imagery with current observations, producing real-time assessments for applications such as disaster mapping and building damage evaluation. These capabilities provide critical support for emergency response and resource allocation in time-sensitive scenarios.

Early CNN-based RSCD methods [20, 21, 27, 94, 108] offer strong efficiency and deployment friendliness, but are inherently limited by local receptive fields, making it difficult to aggregate image-level global context and often leading to pronounced pseudo changes caused by illumination, seasonal, and radiometric

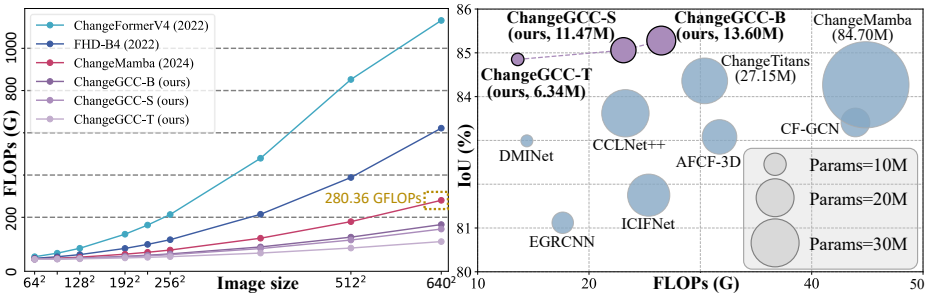

Fig. 1: Efficiency comparison across representative RSCD models. The left panel shows the scaling of computational cost (FLOPs) with input resolution, demonstrating that ChangeGCC grows more gracefully than attention and SSM-based counterparts. The right panel illustrates the accuracy-efficiency trade-off on LEVIR-CD.

variations. In contrast, Vision Transformers [13, 24, 33, 52, 53, 99] leverage token-to-token self-attention to model long-range dependencies and effectively suppress such spurious responses, yet their quadratic complexity results in rapidly growing computational cost as image resolution increases. Linear attention [4, 42, 61–63] and state-space models [32, 35, 51, 110] aim to alleviate this issue by reducing asymptotic complexity. However, these models are primarily designed for sequential processing and require reshaping 2D feature maps into token sequences, making native image-level global reasoning indirect. As a result, their ability to capture rich spatial context in dense prediction tasks often relies on additional architectural components, which substantially increases practical complexity. A representative example is ChangeMamba [16], where visual state-space models alone are insufficient to capture global spatio-temporal dependencies for RSCD, necessitating a heavy Spatio-Temporal Relationship Module (STRM) to compensate for this limitation. While improving change discrimination, this auxiliary design substantially inflates model complexity. As shown in Fig. 1, at a resolution of 640^2, ChangeMamba reaches 84.70M parameters and 280.36G FLOPs, exceeding typical resource budgets for UAV and satellite platforms. Consequently, despite linear asymptotic complexity, such approaches struggle to achieve a favorable accuracy-efficiency balance in practical change detection deployments.

Motivated by these challenges, recent studies have begun to enrich convolutional architectures with global context through large-kernel designs [22], dynamic filtering [18,78,97], and hybrid convolution-attention mechanisms [79,109]. These efforts demonstrate that convolutional models can effectively capture long-range dependencies and driven progress across generic vision tasks, suggesting a promising paradigm for combining global awareness with convolutional efficiency.

Building on this emerging paradigm, we move beyond attention and propose **ChangeGCC**, a fully convolutional, linear-time framework for RSCD with efficient global context. ChangeGCC comprises two components: (a) a **hierarchical backbone** based on **G**lobally **C**onditioned **C**onvolution (GCC), which builds multi-resolution representations with convolutional global routing, and

(b) a **Dual-Temporal Fusion** (DTF) module that explicitly models cross-scale and cross-temporal interactions. DTF follows a two-stage design, featuring (b1) intra-frame cross-scale aggregation to integrate complementary semantics across resolutions, and (b2) inter-frame cross-temporal interaction to mutually condition the two timestamps, highlighting genuine structural changes while suppressing pseudo variations. This design reconciles global awareness with convolutional efficiency, positioning ChangeGCC favorably along the accuracy-efficiency spectrum. As shown in Fig. 1, ChangeGCC advances the trade-off frontier of RSCD, enabling high-performance inference under constrained computational budgets.

Our main contributions are summarized as follows:

- We propose ChangeGCC, a *fully convolutional*, *linear-time* RSCD framework that enables convolutional global routing via Globally Conditioned Convolution, effectively reconciling global contextual modeling with efficiency.
- We design a Dual-Temporal Fusion (DTF) module that performs cross-scale aggregation and cross-temporal interaction, enhancing semantic consistency while suppressing pseudo changes from illumination and seasonal variations.
- We conduct extensive experiments on five optical and SAR change detection benchmarks, demonstrating that ChangeGCC achieves state-of-the-art accuracy while substantially reducing computational complexity, establishing a superior accuracy-efficiency trade-off for resource-constrained deployment.

2 Related Works

CNN-based Change Detection. CNN-based methods constitute a long standing paradigm for RSCD, leveraging hierarchical convolutional features for efficient pixel-level change modeling. FC-Siam [21] introduced fully convolutional Siamese architectures with early fusion or feature-level differencing for dual-temporal analysis. SNUNet [27] combined Siamese structures with densely connected U-Net designs to enhance multi-scale feature fusion. ESCNet [102] incorporated differentiable superpixels to strengthen boundary modeling, while TinyCD [20] and RDP-Net [15] focused on lightweight designs for deployment-oriented efficiency. Other works explored instance-level augmentation [12], spatial-spectral collaboration [44], early-fusion supervision [86], and 3D convolutions [94] to jointly capture spatial and temporal cues. While CNN-based approaches offer strong efficiency [6,7,9,31,58,91], their limited receptive fields hinder image-level global reasoning, often leading to pronounced pseudo changes in complex scenes.

Transformer-based Change Detection. Transformer-based methods [59,106] address the locality of CNNs by introducing self-attention for long-range dependency modeling. BIT [13] projected dual-temporal images into token space for global context reasoning before decoding to pixel predictions. ChangeFormer [3] integrated Transformers with Siamese encoders to enhance cross-temporal alignment and multi-scale decoding. DSTCNet [72] factorized spatial and temporal attention to separately model global context and temporal correspondence. FHD [60] decouples temporal feature extraction and applies hierarchical difference fusion to enhance change discrimination. Changer [26] introduced feature interac-

tion layers for efficient dual-temporal exchange. MCAT [17] incorporated multi-scale attention mechanisms to emphasize change-relevant representations across resolutions, while subsequent works explored multimodal guidance [5, 39, 88], sparse attention [55], deformable attention [98], explicit alignment [95, 96, 100], and probabilistic formulations [105]. These attention-based approaches improve global modeling and change discrimination, but their quadratic complexity limits scalability to high-resolution imagery and resource-constrained platforms.

Linear Attention and State-Space Models. To alleviate the quadratic cost of Transformers, recent works explored linear attention and State-Space Models (SSMs) for efficient global dependency modeling. ChangeMamba [16] introduced visual SSMs into RSCD with task-specific spatio-temporal decoders. M-CD [57] adopted siamese Mamba encoders with multi-scale difference fusion. CD-Mamba [101] embedded convolutions into SSMs to enhance local details, while RS-Mamba [104] proposed omnidirectional selective scanning for large-scale prediction. ChangeTitans [93] incorporated neural memory for long-range context, and ChangeRWKV [90] adapted recurrent-weighted key-value for linear-time bi-temporal reasoning. Despite reduced complexity, these methods typically require 2D-to-sequence reshaping and auxiliary spatio-temporal modules, incurring notable overhead. In contrast, ChangeGCC maintains native 2D convolutions with efficient global context, providing a simpler and deployment-friendly alternative.

3 Method

Given a pair of remote sensing images $(I_1, I_2) \in \mathbb{R}^{3\times H\times W}$, ChangeGCC predicts a dense change map $Y \in \{0,1\}^{H\times W}$. As illustrated in Fig. 2, the framework consists of three main components: (a) a hierarchical backbone based on Globally Conditioned Convolution (GCC), which encodes each timestamp into a four-level feature pyramid $\{F_{i,j}\}_{j=1}^{4}$ with convolutional global routing, where $i \in \{1,2\}$ denotes the two timestamps and $F_{i,j} \in \mathbb{R}^{C_j\times(H/2^j)\times(W/2^j)}$ denotes the feature map at scale j; (b) a Dual-Temporal Fusion (DTF) module that performs cross-scale and cross-temporal interactions, where intra-frame cross-scale aggregation integrates multi-scale features $\{F_{i,j}\}_{j=1}^{4}$ into scale-consistent representations $F'_{i,j}$ for each timestamp i, followed by inter-frame cross-temporal interaction that fuses $\{F'_{i,j}\}_{i=1}^{2}$ at each scale j to produce temporally-aware features F_j; and (c) a lightweight decoder that aggregates the fused multi-scale features $\{F_j\}_{j=1}^{4}$ and upsamples them to generate the final change probability map $\hat{Y} \in (0,1)^{H\times W}$.

3.1 Hierarchical Backbone

To enable global contextual modeling in a linear-time framework, we draw inspiration from [18, 78, 97] and introduce Globally Conditioned Convolution (GCC) based on content-adaptive kernel generation. Unlike standard convolutions with fixed spatial kernels, GCC dynamically yields globally conditioned kernels, enabling local aggregation to incorporate holistic context without explicit token-to-token interactions, avoiding the quadratic complexity of self-attention.

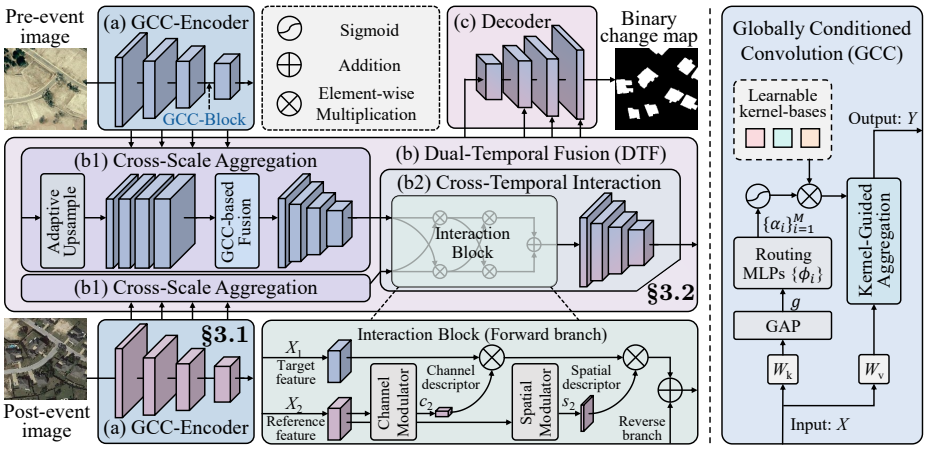

Fig. 2: Overall architecture of ChangeGCC. The model consists of three main components: (a) a GCC-based hierarchical encoder that extracts multi-level features, (b) a Dual-Temporal Fusion (DTF) module that integrates cross-scale and temporal cues, and (c) a lightweight decoder that generates the change mask. DTF is further decomposed into (b1) a Cross-Scale Aggregation module for intra-frame multi-level fusion and (b2) a Cross-Temporal Interaction module for bi-temporal feature conditioning.

Globally Conditioned Convolution. Given an $X \in \mathbb{R}^{C\times h\times w}$, GCC performs input-adaptive convolution by dynamically routing among a set of learnable spatial kernels under global context guidance. First, a compact *global descriptor* is obtained from a projected representation $g = \texttt{GAP}(W_k X)$, where $\texttt{GAP}(\cdot)$ is global average pooling. This descriptor encodes scene-level statistics and is transformed through lightweight generators to produce *routing coefficients* $\alpha_i(g) = \sigma(\phi_i(g))$, where $\phi_i(\cdot)$ are learnable mappings and $\sigma(\cdot)$ denotes the sigmoid activation. These coefficients are used to assemble an adaptive convolution kernel K from a set of *learnable kernel bases* $\{K_i\}_{i=1}^{M}$ via $K = \sum_{i=1}^{M} \alpha_i(g) K_i$, resulting in a globally conditioned kernel that remains compatible with standard local aggregation:

$$Y(p) = \sum_{q\in\mathcal{N}(p)} K(p-q)\cdot V(q), \quad V = W_v X. \tag{1}$$

Although aggregation is locally performed over $\mathcal{N}(p)$, the kernel K is globally conditioned on the entire feature map, enabling convolutional global routing without explicit token-to-token attention while retaining linear complexity.

Hierarchical Architecture. As shown in Fig. 2(a), building upon GCC blocks, we construct a four-stage hierarchical backbone that produces a multi-scale feature pyramid $\{F_{i,j}\}_{j=1}^{4}$ from each input image. Unlike ViT-like backbones [24, 25, 81] that reshape 2D features into token sequences, our hierarchical encoder design maintains spatially-structured feature maps across scales, facilitating precise localization and multi-scale reasoning for dense RSCD task.

3.2 Dual-Temporal Fusion

Given the multi-scale feature pyramids $\{\{F_{i,j}\}_{j=1}^{4}\}_{i=1}^{2}$ extracted from the two timestamps, we design a Dual-Temporal Fusion (DTF) module to explicitly model both cross-scale and cross-temporal interactions. As shown in Fig 2(b), DTF follows a two-stage pipeline: (b1) intra-frame cross-scale aggregation integrates complementary semantics across resolutions per timestamp, and (b2) inter-frame cross-temporal interaction mutually conditions the two timestamps to highlight genuine structural changes while suppressing pseudo variations.

Intra-frame Cross-Scale Aggregation. For each timestamp i, DTF first aggregates the multi-scale features $\{F_{i,j}\}_{j=1}^{4}$ to construct scale-consistent representations. Concretely, all feature maps are upsampled to the highest spatial resolution and fused via a GCC block introduced in Sec. 3.1:

$$\tilde{F}_i = \texttt{GCC}(\texttt{Concat}(\{\texttt{Up}(F_{i,j})\}_{j=1}^{4})), \tag{2}$$

where $\texttt{Up}(\cdot)$ denotes bilinear upsampling and $\texttt{Concat}(\cdot)$ concatenates features along the channel dimension. The fused feature $\tilde{F}_i$ is then projected back to each scale to obtain refined representations $F'_{i,j} = \texttt{Down}_j(F_i)$, where $\texttt{Down}_j(\cdot)$ denotes scale-specific downsampling. This step enables information exchange across resolutions, enhancing semantic consistency while preserving spatial details.

Inter-frame Cross-Temporal Interaction. Given the scale-aligned features $\{F'_{i,j}\}_{i=1}^{2}$, DTF performs cross-temporal interaction at each scale j. Inspired by cross-attention but implemented in a convolutional manner, we adopt a symmetric interaction scheme where features from one timestamp modulate the other. Specifically, let $X_1 = F'_{1,j}$ and $X_2 = F'_{2,j}$. Channel descriptors are computed via global average and max pooling, yielding cross-temporal channel gating as:

$$c_i = \sigma(\phi_i(\texttt{GAP}(X_i)) + \phi_i(\texttt{GMP}(X_i))), \quad \tilde{X}_1 = X_1 \odot c_2, \quad \tilde{X}_2 = X_2 \odot c_1, \tag{3}$$

where $\phi_i(\cdot)$ denotes a lightweight bottleneck network and $\sigma(\cdot)$ is the sigmoid function. Spatial descriptors are then derived through channel-wise pooling and a small convolutional predictor, followed by symmetric spatial conditioning:

$$s_i = \sigma(\psi_i(\texttt{Concat}(\texttt{Mean}_c(\tilde{X}_i), \texttt{Max}_c(\tilde{X}_i)))), \quad \hat{X}_1 = \tilde{X}_1 \odot s_2, \quad \hat{X}_2 = \tilde{X}_2 \odot s_1, \tag{4}$$

where $\psi_i(\cdot)$ denotes a lightweight spatial gating function. The fused representation at scale j is obtained via symmetric aggregation, $F_j = \hat{X}_1 + \hat{X}_2$, resulting in cross-channel and cross-spatial mutual conditioning across timestamps.

Decoder. The temporally fused multi-scale features $\{F_j\}_{j=1}^{4}$ are finally aggregated by a UNet-like [65] convolutional decoder to produce the change map $\hat{Y}$.

3.3 Loss Function

We train ChangeGCC end-to-end with a hybrid objective that combines pixel-wise binary cross-entropy (BCE) and Dice loss to balance per-pixel supervision

and region-level overlap. Given the predicted change map $\hat{Y} \in (0,1)^{H\times W}$ and the binary ground truth $Y \in \{0,1\}^{H\times W}$, the BCE term is defined as:

$$\mathcal{L}_{\mathrm{BCE}}(\hat{Y}, Y) = -\frac{1}{HW}\sum_{p}\Big(Y(p)\log\hat{Y}(p) + (1-Y(p))\log(1-\hat{Y}(p))\Big). \quad (5)$$

We further adopt the soft Dice coefficient and its corresponding loss:

$$\mathrm{Dice}(\hat{Y}, Y) = \frac{2\sum_p \hat{Y}(p)Y(p) + \epsilon}{\sum_p \hat{Y}(p) + \sum_p Y(p) + \epsilon}, \quad \mathcal{L}_{\mathrm{Dice}}(\hat{Y}, Y) = 1 - \mathrm{Dice}(\hat{Y}, Y) \quad (6)$$

The final training objective is $\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \lambda\mathcal{L}_{\mathrm{Dice}}$, where λ balances the two terms, pixel-wise supervision and region-level overlap, and is set to 1 by default.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate ChangeGCC on five public RSCD benchmarks across diverse scenes, resolutions, and sensing modalities, enabling a comprehensive assessment of its effectiveness and generalization ability. Details are as follows:

- **LEVIR-CD** [14]: A large-scale binary RSCD dataset containing 637 VHR (0.5 m/pixel) Google Earth image pairs of size 1024^2, focusing on building changes in urban scenes with expert-annotated pixel-wise change labels.
- **WHU-CD** [40]: A single pair of large-scale aerial images over Christchurch, New Zealand (2012&2016, 0.2 m/pixel), with an original resolution of 32507×15354, typically cropped into 256^2 patches for training and evaluation.
- **LEVIR-CD+** [66]: An extended version of LEVIR-CD comprising 985 VHR (0.5 m/pixel) image pairs of size 1024^2, providing more challenging samples and refined annotations for small-scale and fine-boundary change detection.
- **SYSU-CD** [73]: A large-scale dataset containing 20,000 aerial image pairs of size 256^2 collected in Hong Kong, covering diverse change types and exhibiting appearance variations caused by illumination and seasonal effects.
- **SAR-CD** [1]: A SAR-based dataset containing 10,000 Sentinel-1 SAR image pairs of size 512^2, with synthetic changes generated from single-temporal imagery, characterized by speckle noise and radiometric inconsistencies.

Implementation Details. All datasets are preprocessed following the OpenCD [45] protocol, where bi-temporal images are cropped into patches of size 256^2 for training and inference. All experiments are conducted on 4 NVIDIA GeForce RTX 3090 GPUs, with a batch size of 8 per GPU. We follow the training configuration of [92,94], optimizing the proposed ChangeGCC for a total of 200 epochs using Adam [43]. The initial learning rate is set to 1×10^{-5}, with a weight decay of 1×10^{-4} and momentum of 0.999. Gradient clipping is applied with a margin of 0.5. A cosine learning rate schedule is adopted, together with an initial 20

epochs warmup phase and a warmup multiplier of 10. Data augmentation adopts a two-stage design: random shared geometric transformations are first applied to preserve spatial alignment, followed by independent photometric perturbations on each timestamp to simulate pseudo changes and improve robustness.

To balance accuracy and efficiency, we further instantiate ChangeGCC at three model scales with different encoder depths and embedding dimensions $\{C_j\}_{j=1}^{4}$: ChangeGCC-T, ChangeGCC-S, and ChangeGCC-B, corresponding to tiny, small, and base configurations. The specific architectures are as follows:

- **ChangeGCC-T** uses encoder depths of [2,2,4,2] and $\{C_j\}$=[32,64,128,240].
- **ChangeGCC-S** uses encoder depths of [2,2,7,2] and $\{C_j\}$=[40,80,160,320].
- **ChangeGCC-B** uses encoder depths of [2,3,10,3] and $\{C_j\}$=[40,80,160,320].

Evaluation Metrics. We evaluate performance using standard pixel-level metrics, including Precision (P), Recall (R), F1-score, and Intersection over Union (IoU). To further assess boundary preservation and the localization accuracy of small-scale changes, we additionally employ three boundary-aware metrics: Boundary F1-score (BF1), Trimap-based mIoU, and Hausdorff distance (H). Specifically, BF1 is computed from boundary precision (P_b) and boundary recall (R_b), Trimap-mIoU is evaluated within a narrow trimap region using TP_t, FP_t, and FN_t, while Hausdorff distance measures the maximum discrepancy between the predicted boundary set P and the ground-truth boundary set G:

$$\text{BF1} = \frac{2P_bR_b}{P_b + R_b}, \quad \text{Trimap-mIoU} = \frac{TP_t}{TP_t + FP_t + FN_t}, \tag{7}$$

$$H(P, G) = \max\left(\sup_{p\in P}\inf_{g\in G}|p-g|, \sup_{g\in G}\inf_{p\in P}|g-p|\right). \tag{8}$$

These metrics provide a fine-grained assessment of boundary quality, structural consistency, and localization accuracy beyond region-level overlap measures.

4.2 Comparison with State-of-the-Art Methods

We compare ChangeGCC against representative state-of-the-art (SOTA) RSCD methods across all five benchmarks under identical experimental settings. All competing approaches are evaluated using their official implementations or reproduced following the authors' configurations to ensure strictly fair comparison.
Performance on Standard Optical Benchmarks. Quantitative comparisons on the LEVIR-CD and WHU-CD datasets are reported in Tab. 1. Across both benchmarks, ChangeGCC achieves competitive performance while maintaining significantly lower computational cost. In particular, ChangeGCC-B attains an IoU of 85.28% and F1-score of 92.06% on LEVIR-CD, and 89.95% / 94.71% on WHU-CD, outperforming strong recent baselines based on linear-attention such as ChangeMamba [16] (84.27% IoU @ 44.86G FLOPs on LEVIR-CD) and ChangeTitans [92] (84.36% IoU @ 30.39G FLOPs). Notably, even the lightweight variants, ChangeGCC-T, deliver strong accuracy (84.85% IoU @

Table 1: Comparison results on the LEVIR-CD [14] and WHU-CD [40] datasets. The top three results are marked in red (1st), blue (2nd), and green (3rd).

Method	Year	Params (M)	FLOPs (G)	LEVIR-CD				WHU-CD			
				IoU	F1	P	R	IoU	F1	P	R
FC-Siam-Diff [21]	2018	4.38	1.35	75.92	86.31	89.53	83.31	41.66	58.81	47.33	77.66
FC-Siam-Conc [21]	2018	4.99	1.55	71.96	83.69	91.99	76.77	49.95	66.63	60.88	73.58
SNUNet [27]	2021	12.03	27.44	78.83	88.16	89.18	87.17	71.67	83.50	85.60	81.49
BIT [13]	2021	3.55	4.35	80.68	89.31	89.24	89.37	72.39	83.98	86.64	81.48
SADL_CD [11]	2022	15.09	7.83	79.75	88.74	90.91	86.66	81.78	89.98	93.10	87.05
MSPSNet [34]	2022	2.21	14.17	81.36	89.18	91.38	87.08	79.04	88.29	89.95	86.70
EGRCNN [2]	2022	9.64	17.64	81.12	89.59	92.83	86.46	75.03	84.86	90.23	81.09
ICIFNet [29]	2022	23.82	25.36	81.75	89.96	91.32	88.64	79.24	88.32	92.98	85.56
ChangeFormer [3]	2022	41.02	202.80	82.48	90.40	92.05	88.80	71.91	83.66	85.49	81.90
DMINet [28]	2023	6.24	14.42	82.99	90.71	92.52	89.95	79.68	88.69	93.84	86.25
WNet [77]	2023	43.07	19.20	82.93	90.67	91.16	90.18	83.91	91.25	92.37	90.15
AFCF-3D [94]	2023	17.54	31.72	83.08	90.76	91.35	90.17	87.93	93.58	93.47	93.69
VcT [41]	2023	3.57	10.64	81.89	90.04	92.57	87.65	81.12	89.58	89.39	89.77
CCLNet++ [74]	2023	28.78	23.27	83.62	91.08	92.31	89.88	82.32	90.31	89.83	90.78
S^2CD [83]	2024	3.22	9.26	82.78	90.58	92.12	89.08	82.13	90.19	92.45	88.74
CF-GCN [84]	2024	13.58	43.93	83.41	90.96	91.75	90.18	84.90	91.83	94.81	89.04
SEIFNet [38]	2024	8.37	27.91	83.40	90.95	92.49	89.46	76.04	86.39	87.01	85.77
CBSASNet [36]	2024	5.76	31.64	84.14	91.39	92.47	90.33	86.08	92.52	93.93	91.15
ChangeMamba [16]	2024	84.70	44.86	84.27	91.37	93.87	89.00	88.02	93.63	94.22	93.05
ChangeTitans [92]	2025	27.15	30.39	84.36	91.52	92.49	90.56	88.56	94.10	94.33	93.87
AMDANet [75]	2025	25.83	77.23	82.34	90.32	92.45	88.28	80.68	89.30	88.77	89.84
S^2ENet [85]	2025	24.92	17.59	84.61	91.67	92.03	91.30	89.33	94.36	95.67	93.09
MEDS-CD [47]	2026	46.39	276.55	84.77	90.75	91.43	90.08	88.26	93.76	95.58	92.01
ChangeGCC-T (ours)	-	6.34	13.60	84.85	91.80	92.40	91.22	89.52	94.47	95.34	93.62
ChangeGCC-S (ours)	-	11.47	23.09	85.06	91.92	92.76	91.11	89.48	94.56	95.11	94.02
ChangeGCC-B (ours)	-	13.60	26.49	85.28	92.06	92.87	91.25	89.95	94.71	96.17	93.30

13.60G FLOPs) with substantially reduced parameters and FLOPs, demonstrating favorable accuracy-efficiency trade-offs. Compared with recent large-scale models like ChangeTitans, ChangeGCC achieves comparable or better performance with markedly lower computational overhead, highlighting the effectiveness of globally conditioned convolutions for efficient global context modeling.

Robustness to Appearance Variations and Cross-Category Changes. The quantitative comparisons on LEVIR-CD+ and SYSU-CD are reported in Tab. 2 and Tab. 3, respectively. LEVIR-CD+ emphasizes fine-grained boundaries and small-scale changes, while SYSU-CD involves diverse cross-category land-cover transitions with pronounced appearance variations. On LEVIR-CD+ (Tab. 2), ChangeGCC-B achieves the best IoU of 75.92%, surpassing recent strong methods such as ChangeTitans [92], ChangeCLIP [23], and SA-CDNet [82]. Notably, ChangeGCC-S and ChangeGCC-T also deliver competitive performance (74.77% and 74.06%, respectively) with substantially fewer parameters and FLOPs than Transformer- and SSM-based counterparts, demonstrating favorable accuracy-efficiency trade-offs. On the more challenging SYSU-CD benchmark (Tab. 3), ChangeGCC-B attains the highest IoU of 71.54%, outperforming ChangeTitans [92], MGCR-Net [80], and ChangeDA [54]. Since the SYSU-CD dataset contains cross-category semantic changes beyond building-only scenar-

Table 2: Comparison results on the LEVIR-CD+ [66] dataset. All metrics are reported as percentages (%), with the highest value highlighted in **bold**.

Method	Year	Params (M)	FLOPs (G)	IoU
FC-Siam-Diff [21]	2018	4.38	1.35	57.44
FC-Siam-Conc [21]	2018	4.99	1.55	58.07
STANet [14]	2020	16.93	6.58	65.66
SNUNet [27]	2021	12.03	27.44	67.11
BIT [13]	2021	3.55	4.35	70.64
MFPNet [87]	2021	60.06	107.73	73.22
ICIFNet [29]	2022	23.82	25.36	71.61
TransUNetCD [46]	2022	28.37	244.54	71.86
SwinSUNet [99]	2022	39.28	43.50	74.82
MSCANet [49]	2022	16.59	14.68	72.54
FCCDN [19]	2022	6.31	12.52	72.76
CTDFormer [103]	2023	3.85	303.77	67.09
AFCF-3D [94]	2023	17.54	31.72	70.53
BiFA [100]	2024	5.58	53.00	72.35
ChangeCLIP [23]	2024	-	-	75.63
SAM2-CD [64]	2025	72.53	567.81	68.94
ChangeTitans [92]	2025	27.15	30.39	75.84
ConvFormer-CD/48 [89]	2025	37.72	5.14	74.37
SPMNet [82]	2025	17.33	6.94	71.01
SA-CDNet [82]	2025	77.08	12.03	73.04
ChangeGCC-T (ours)	-	6.34	13.60	74.06
ChangeGCC-S (ours)	-	11.47	23.09	74.77
ChangeGCC-B (ours)	-	13.60	26.49	**75.92**

Table 3: Comparison results on the SYSU-CD [73] dataset. All metrics are reported as percentages (%), with the highest value highlighted in **bold**.

Method	Year	Params (M)	FLOPs (G)	IoU
FC-Siam-Diff [21]	2018	4.38	1.35	56.96
FC-Siam-Conc [21]	2018	4.99	1.55	61.75
STANet [14]	2020	16.93	6.58	63.09
SNUNet [27]	2021	12.03	27.44	66.35
BIT [13]	2021	3.55	4.35	61.39
ICIFNet [29]	2022	23.82	25.36	68.12
TransUNetCD [46]	2022	28.37	244.54	66.79
SwinSUNet [99]	2022	39.28	43.50	68.89
CTDFormer [103]	2023	3.85	303.77	64.04
WNet [77]	2023	43.07	19.20	67.55
SEIFNet [38]	2024	8.37	27.91	69.96
MutSimNet [50]	2024	45.44	58.10	70.27
STENet [56]	2024	10.95	14.73	69.92
AMDANet [75]	2025	25.83	77.23	67.82
ChangeDA [54]	2025	30.25	23.49	70.56
ChangeTitans [92]	2025	27.15	30.39	71.18
FILFBCD [37]	2026	638.60	183.89	69.16
BSCNet [107]	2026	45.30	39.47	70.29
MEDS-CD [47]	2026	46.39	276.55	68.80
MGCR-Net [80]	2026	77.71	19.31	70.80
ChangeGCC-T (ours)	-	6.34	13.60	70.35
ChangeGCC-S (ours)	-	11.47	23.09	70.98
ChangeGCC-B (ours)	-	13.60	26.49	**71.54**

ios, these results indicate that our ChangeGCC generalizes effectively to diverse land-cover transformations. Overall, the consistent gains across both datasets validate the robustness of ChangeGCC under fine-grained variations and heterogeneous change patterns, while maintaining strong computational efficiency.

Qualitative Analysis. We present qualitative comparisons on the LEVIR-CD (building scenes) and SYSU-CD (non-building scenes) datasets in Fig. 3, against BIT [13], ChangeMamba [16], and AFCF-3D [94]. These examples highlight typical challenges in RSCD, including dense object layouts, irregular boundaries, and appearance-induced pseudo changes. On LEVIR-CD, ChangeGCC generates cleaner predictions with sharper boundaries and more complete building footprints, especially in dense areas and around irregular structures, where competing methods often suffer from fragmentation or blurred edges. On the more challenging SYSU-CD, ChangeGCC shows stronger robustness to appearance variations and background clutter, producing fewer false positives and more accurate localization of elongated objects. Across both datasets, ChangeGCC preserves object geometry and boundary details, which aligns with its superior BF1, Trimap-mIoU, and Hausdorff distance, validating the effectiveness of our dual-temporal fusion and GCC design for structural modeling.

Cross-Modality Robustness on SAR Imagery. We first report quantitative comparisons on SAR-CD in Tab. 4. SAR-CD is particularly challenging as changes are synthetically generated [8, 10], requiring models to directly capture

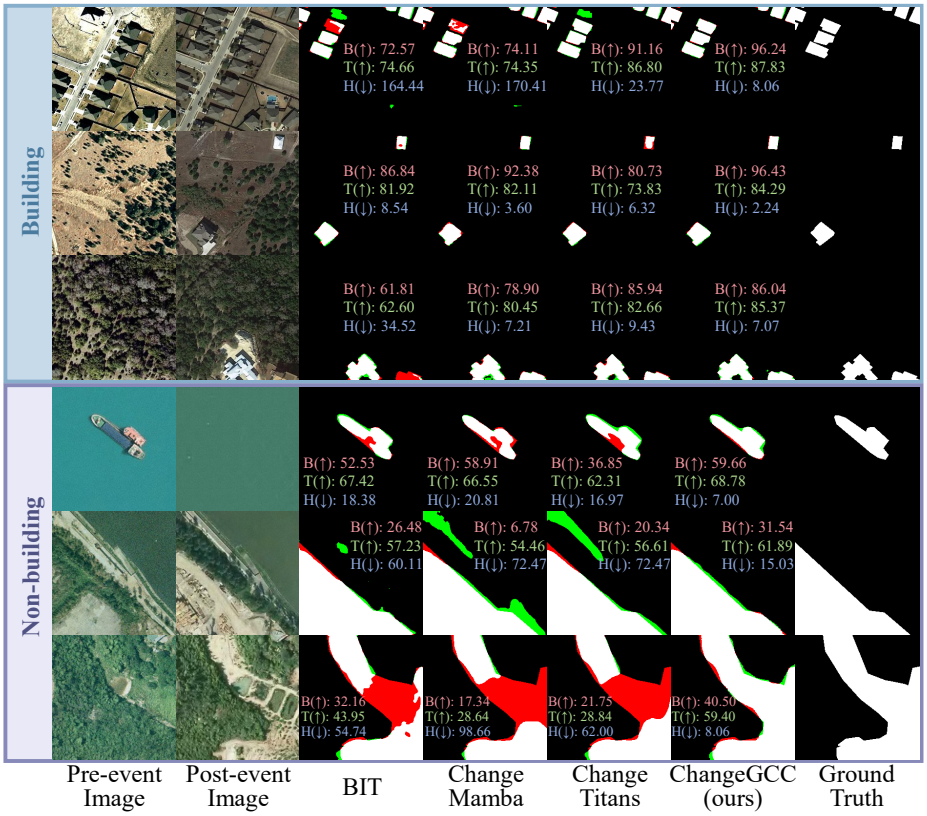

Fig. 3: Qualitative comparisons on the LEVIR-CD [14] (building) and SYSU-CD [73] (non-building) datasets. Predicted outputs are color-coded as follows: white for true positives (TP), black for true negatives (TN), green for false positives (FP), and red for false negatives (FN). In the figure, B represents Boundary F1-score, T represents Trimap-based mIoU, and H represents Hausdorff distance.

subtle difference patterns under strong speckle noise and radiometric inconsistencies. ChangeGCC achieves the best performance across all model scales, with ChangeGCC-B reaching 97.21% IoU, outperforming both CNN-based and recent linear-attention methods while using significantly fewer parameters and FLOPs than ChangeMamba. Notably, despite its lightweight design, ChangeGCC consistently surpasses ChangeTitans and M-CD, demonstrating strong robustness under SAR noise. We further provide qualitative comparisons against BIT and AFCF-3D in Fig. 4. Competing methods tend to introduce fragmented responses or miss fine structures in noisy regions, whereas ChangeGCC produces cleaner change maps with more coherent object geometry and fewer false alarms. These results indicate that ChangeGCC generalizes well beyond optical imagery and maintains strong performance on SAR data, validating its ability to directly model change cues under cross-modality and noise-dominated conditions.

Table 4: Comparison results on the SAR-CD [1] dataset. All metrics are reported as percentages (%), with the highest value highlighted in **bold**. For fair comparison, all methods are trained under the same training schedule as [92].

Method	Year	Params (M)	FLOPs (G)	IoU
FC-SiamDiff [21]	2018	4.38	1.35	86.07
FC-SiamConc [21]	2018	4.99	1.55	93.39
STANet [14]	2020	16.93	6.58	92.87
SNUNet [27]	2021	12.03	27.44	93.52
BIT [13]	2021	3.55	4.35	93.48
MFPNet [87]	2021	60.06	107.73	97.09
MSCANet [49]	2022	16.59	14.68	94.43
FHD [60]	2022	11.83	10.43	95.56
FCS [19]	2022	3.03	9.59	94.77
DED [19]	2022	3.10	9.90	94.94
FCCDN [19]	2022	6.31	12.52	95.15
SwinSUNet [99]	2022	39.28	43.50	94.61
ChangeFormer [3]	2022	41.02	202.80	96.47
AFCF-3D [94]	2023	17.54	31.72	93.72
ChangeMamba [16]	2024	84.70	44.86	90.40
ChangeTitans [92]	2025	27.15	30.39	95.63
M-CD [57]	2025	69.80	29.58	95.76
ChangeGCC-T (ours)	-	6.34	13.60	96.53
ChangeGCC-S (ours)	-	11.47	23.09	96.91
ChangeGCC-B (ours)	-	13.60	26.49	**97.21**

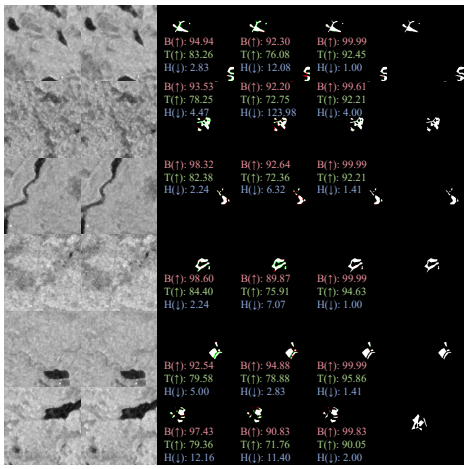

Fig. 4: Qualitative results on SAR-CD [1]. Predicted outputs are color-coded as follows: white for true positives (TP), black for true negatives (TN), green for false positives (FP), and red for false negatives (FN).

Robustness to Label Noise. Label noise [30, 67–71, 76] refers to inaccurate or missing annotations in ground-truth labels. In RSCD, such noise typically manifests as incomplete or misaligned masks, which commonly occurs in large-scale RSCD datasets such as LEVIR-CD. As shown in Fig. 5, ChangeGCC remains robust under noisy supervision and produces consistent predictions when annotations contain obvious errors. This property is particularly important for practical RSCD deployment, where clean annotations are rarely available at scale.

4.3 Ablation Study

We conduct ablation studies on LEVIR-CD using ChangeGCC-B as the default setting to systematically evaluate the contribution of each architectural component. As shown in Tab. 5, key modules are removed or replaced while keeping all other configurations fixed. Performance is measured by IoU and F1-score.

Kernel Strategy (*cf.* ❶ ❷). We first investigate the impact of different kernel adaptation strategies in the vision encoder. Replacing the proposed GCC with standard fixed convolution kernels leads to clear performance degradation. Specifically, using 3×3 or 7×7 convolutions reduces IoU from 85.28% to 81.80% and 83.24%, respectively. In contrast, GCC achieves the best result with 85.28% IoU and 92.06% F1-score, demonstrating that globally conditioned kernels provide more effective contextual modeling than simply enlarging receptive fields.

Table 5: Ablation study on different architectural components. Rows highlighted in purple indicate the default configuration.

Experiment	Variant #	Method	IoU	F1
Component				
Kernel Strategy	❶	3×3 Conv	81.80	89.99
	❷	7×7 Conv	83.24	90.86
	-	**GCC (ours)**	**85.28**	**92.06**
Cross-Scale Aggregation	❸	*w/o*	84.53	91.58
	❹	FPN [48]	84.96	91.90
	-	**ours**	**85.28**	**92.06**
Cross-Temporal Interaction	❺	Early Fusion	80.55	89.23
	❻	SiamDiff [21]	82.06	90.62
	❼	SiamConc [21]	82.44	90.79
	❽	FHD [60]	84.93	91.73
	❾	STRM [16]	84.72	91.34
	-	**ours**	**85.28**	**92.06**
Hyperparameter				
Balance Factor λ	①	0	84.97	91.90
	②	0.5	85.02	91.88
	-	**1.0 (ours)**	**85.28**	**92.06**
	③	2.0	84.60	91.58

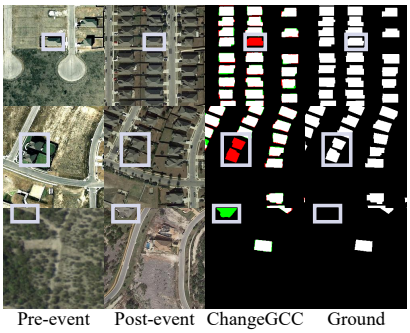

Fig. 5: Qualitative comparison on the LEVIR-CD [14] dataset under label noise. Here, label noise refers to incomplete or incorrect annotations in the training set, as highlighted in the purple boxed regions.

Cross-Scale Aggregation (*cf.* ❸ ❹). We next ablate the intra-frame cross-scale aggregation module. Removing this component or replacing it with a standard FPN [48] structure both lead to noticeable performance degradation. Although FPN provides a top-down feature fusion pathway, it performs scale alignment in a progressively feed-forward and scale-independent manner. In contrast, our design explicitly models cross-scale interactions via attentive aggregation, allowing features at different resolutions to mutually condition each other. This leads to more coherent hierarchical multi-scale representations and achieves the highest accuracy, highlighting the importance of explicit cross-scale reasoning.

Cross-Temporal Interaction (*cf.* ❺ ❻ ❼ ❽ ❾) To evaluate different temporal fusion strategies, we replace our dual-temporal interaction with several common alternatives, including early fusion, Siamese difference and concatenation [21], FHD [60], and the spatio-temporal relationship module (STRM) from ChangeMamba [16]. These variants consistently underperform our design, with IoU ranging from 80.55% to 84.93%. Early fusion and Siamese-based strategies rely on static feature combination or simple difference operations, which lack explicit cross-temporal conditioning. Although FHD and STRM introduce more structured temporal modeling, they do not perform symmetric mutual modulation between timestamps. In contrast, our cross-temporal interaction enables bidirectional conditioning at channel and spatial levels, allowing each timestamp to adaptively filter and refine the other. This mutual interaction leads to more precise temporal correspondence modeling and achieves the best performance.

Balance Factor λ (*cf.* ① ② ③). Finally, we study the influence of the balance factor λ between BCE and Dice losses. Setting $\lambda = 1$ yields the best results, while both smaller and larger values lead to slight performance drops. This suggests

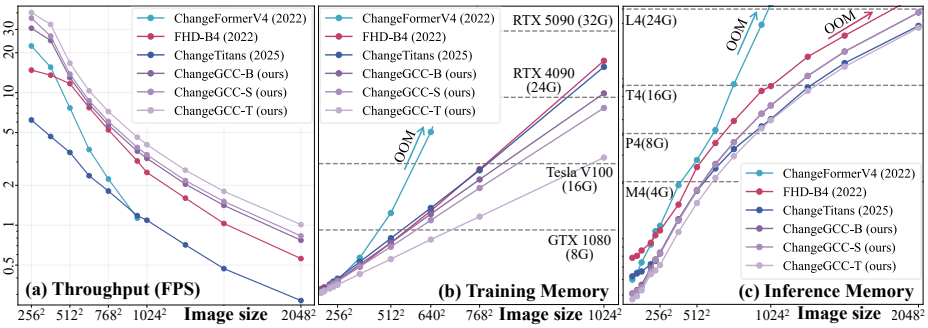

Fig. 6: Efficiency comparison under varying input resolutions. Subfigure (a) reports inference throughput (FPS) measured on an NVIDIA GeForce RTX 3090 (batch size = 1, AMP disabled, after 5 warm-up runs and averaged over the subsequent 10 runs), while (b) and (c) show training and inference memory usage, respectively.

that a balanced combination of pixel-wise supervision and region-level overlap provides stable optimization, and we therefore adopt $\lambda = 1$ as the default setting.

4.4 Efficiency Analysis

Beyond accuracy, we further analyze the computational efficiency of ChangeGCC from both resolution scalability and overall architectural cost perspectives.

Runtime & Memory Scalability. We compare ChangeGCC with representative RSCD models, including ChangeFormer [3], FHD [60], and ChangeTitans [92], in terms of throughput and memory usage under varying input resolutions. As shown in Fig. 6, ChangeGCC consistently requires less memory while achieving higher throughput. Although ChangeTitans exhibits comparable inference memory, its complex neural memory design leads to substantially lower throughput. These results indicate that ChangeGCC maintains stable efficiency under high-resolution scenarios, making it suitable for large-scale applications.

Architectural Efficiency Breakdown. To better understand the allocation of computational cost, Tab. 6 breaks down the parameters and FLOPs of the proposed ChangeGCC into four components: encoder, cross-scale aggregation, cross-temporal interaction, and decoder. Across all model scales, most parameters are concentrated in the encoder and cross-temporal interaction modules, while cross-scale aggregation and the decoder introduce only minimal overhead. In particular, cross-scale aggregation accounts for less than 2% of the total parameters, yet brings notable performance improvements, indicating a favorable efficiency-effectiveness trade-off. From a FLOPs perspective, the encoder dominates the computational cost, followed by cross-temporal interaction, whereas the decoder contributes negligibly. Both global modeling components operate within a moderate budget, demonstrating that effective global reasoning can be achieved without excessive architectural cost. As model capacity increases from Tiny to Base, the relative cost distribution remains consistent, suggesting

Table 6: Architectural cost breakdown of ChangeGCC. We report the parameter count and FLOPs contributed by each component across different model scales. Percentages indicate the proportion of each module within the overall architecture.

Model	Metric	Encoder	Cross-Scale Aggregation	Cross-Temporal Interaction	Decoder	Total
ChangeGCC-T	Params(M)	2.85 (44.95%)	0.12 (1.89%)	3.11 (49.05%)	0.26 (4.10%)	6.34
	FLOPs(G)	5.56 (40.88%)	3.93 (28.90%)	3.59 (26.40%)	0.52 (3.82%)	13.60
ChangeGCC-S	Params(M)	5.54 (48.29%)	0.19 (1.66%)	5.33 (46.46%)	0.41 (3.57%)	11.47
	FLOPs(G)	10.51 (45.52%)	6.09 (26.38%)	5.75 (24.90%)	0.74 (3.20%)	23.09
ChangeGCC-B	Params(M)	7.67 (56.40%)	0.19 (1.40%)	5.33 (39.19%)	0.41 (3.01%)	13.60
	FLOPs(G)	13.91 (52.51%)	6.09 (22.99%)	5.75 (21.71%)	0.74 (2.79%)	26.49

that ChangeGCC scales in a structurally balanced manner without introducing computational bottlenecks. This confirms that the proposed design maintains efficiency not only at the system level but also within its internal architecture.

5 Conclusion

This paper presents **ChangeGCC**, a *fully convolutional, linear-time* framework that addresses the long-standing trade-off between pseudo-change suppression and computational efficiency in remote sensing change detection. By introducing Globally Conditioned Convolution, ChangeGCC enables *convolutional global routing* to capture scene-level context without incurring the quadratic complexity of attention. The proposed dual-temporal fusion module further improves semantic consistency across scales while enhancing the modeling of genuine structural changes. Extensive experiments on optical and SAR benchmarks demonstrate that ChangeGCC achieves state-of-the-art performance while maintaining a lightweight and scalable architecture. Ablation studies verify that both globally conditioned feature extraction and dual-temporal interaction are essential to its effectiveness. Despite these advances, practical challenges such as learning under limited supervision and adapting to diverse sensing conditions still remain open. Future work will further explore weakly-supervised learning and extend ChangeGCC toward broader cross-modal and real-time deployment scenarios.

Acknowledgments

This work was supported by the National Defense Science and Technology Industry Bureau Technology Infrastructure Project (JSZL2024606C001).

References

1. Alatalo, J., Sipola, T., Rantonen, M.: Improved difference images for change detection classifiers in sar imagery using deep learning. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)
2. Bai, B., Fu, W., Lu, T., Li, S.: Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021)
3. Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 207–210. IEEE (2022)
4. Behrouz, A., Zhong, P., Mirrokni, V.: Titans: Learning to memorize at test time. arXiv preprint arXiv:2501.00663 (2024)
5. Bernhard, M., Strauß, N., Schubert, M.: Mapformer: Boosting change detection by using pre-change information. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16837–16846 (2023)
6. Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., Yao, Y.: Poly kernel inception network for remote sensing detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27706–27716 (2024)
7. Cai, X., Li, L., Pei, G., Chen, T., Pan, J., Yao, Y., Wang, W.: Beyond frequency: Scoring-driven debiasing for object detection via blueprint-prompted image synthesis. In: International Conference on Learning Representations (2026)
8. Cai, X., Li, L., Pei, G., Chen, T., Pan, J., Yao, Y., Wang, W.: Unbiased object detection beyond frequency with visually prompted image synthesis. In: The Fourteenth International Conference on Learning Representations (2026)
9. Cai, X., Li, L., Pei, G., Sun, Z., Yao, Y., Wang, W.: Pkinet-v2: Towards powerful and efficient poly-kernel remote sensing object detection. arXiv preprint arXiv:2603.16341 (2026)
10. Cai, X., Pei, G., Sun, Z., Yao, Y., Shen, F., Wang, W.: Iris: Bringing real-world priors into diffusion model for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26909–26919 (2026)
11. Chen, H., Li, W., Chen, S., Shi, Z.: Semantic-aware dense representation learning for remote sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2022)
12. Chen, H., Li, W., Shi, Z.: Adversarial instance augmentation for building change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2021)
13. Chen, H., Qi, Z., Shi, Z.: Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (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. Chen, H., Pu, F., Yang, R., Tang, R., Xu, X.: Rdp-net: Region detail preserving network for change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–10 (2022)
16. Chen, H., Song, J., Han, C., Xia, J., Yokoya, N.: Changemamba: Remote sensing change detection with spatiotemporal state space model. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–20 (2024)
17. Chen, H., Xu, T., Chen, Z., Liu, P., Bai, H., Li, J.: Multi-scale change-aware transformer for remote sensing image change detection. In: Proceedings of the ACM International Conference on Multimedia. pp. 2992–3000 (2024)

18. Chen, L., Gu, L., Li, L., Yan, C., Fu, Y.: Frequency dynamic convolution for dense image prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30178–30188 (2025)
19. Chen, P., Zhang, B., Hong, D., Chen, Z., Yang, X., Li, B.: Fccdn: Feature constraint network for vhr image change detection. ISPRS Journal of Photogrammetry and Remote Sensing **187**, 101–119 (2022)
20. Codegoni, A., Lombardi, G., Ferrari, A.: Tinycd: A (not so) deep learning model for change detection. Neural Computing and Applications **35**(11), 8471–8486 (2023)
21. Daudt, R.C., Le Saux, B., Boulch, A.: Fully convolutional siamese networks for change detection. In: IEEE International Conference on Image Processing. pp. 4063–4067. IEEE (2018)
22. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11963–11975 (2022)
23. Dong, S., Wang, L., Du, B., Meng, X.: Changeclip: Remote sensing change detection with multimodal vision-language representation learning. ISPRS Journal of Photogrammetry and Remote Sensing **208**, 53–69 (2024)
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
25. Duan, Y., Wang, W., Chen, Z., Zhu, X., Lu, L., Lu, T., Qiao, Y., Li, H., Dai, J., Wang, W.: Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. In: International Conference on Learning Representations (2025)
26. Fang, S., Li, K., Li, Z.: Changer: Feature interaction is what you need for change detection. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–11 (2023)
27. 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)
28. 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)
29. Feng, Y., Xu, H., Jiang, J., Liu, H., Zheng, J.: 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 **60**, 1–13 (2022)
30. Gao, Y., Pei, G., Sheng, M., Sun, Z., Chen, T., Yao, Y.: Relating cnn-transformer fusion network for remote sensing change detection. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
31. Garnot, V.S.F., Landrieu, L.: Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4872–4881 (2021)
32. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (2024)
33. Gu, H., Pei, G., Sun, Z., Ren, M., Shu, X., Yao, Y., Shen, F.: Medfg-vqa: Low-frequency memory and graph attention for lightweight medical vqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42755–42764 (2026)

34. Guo, Q., Zhang, J., Zhu, S., Zhong, C., Zhang, Y.: Deep multiscale siamese network with parallel convolutional structure and self-attention for change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–12 (2021)
35. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25261–25270 (2025)
36. He, N., Wang, L., Zheng, P., Zhang, C., Li, L.: Cbsasnet: a siamese network based on channel bias split attention for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing (2024)
37. Huang, J., Ji, S., Wang, Y., Xia, M., Yuan, X.: A sam fine-tuning framework with frequency-domain interactive lora for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–19 (2026)
38. 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)
39. Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., et al.: Terramind: Large-scale generative multimodality for earth observation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7383–7394 (2025)
40. 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**, 574–586 (2018)
41. Jiang, B., Wang, Z., Wang, X., Zhang, Z., Chen, L., Wang, X., Luo, B.: Vct: Visual change transformer for remote sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)
42. Jiang, J., Zhang, J., Liu, W., Gao, M., Hu, X., Yan, X., Huang, F., Liu, Y.: Rwkv-unet: Improving unet with long-range cooperation for effective medical image segmentation. arXiv preprint arXiv:2501.08458 (2025)
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
44. Lei, T., Geng, X., Ning, H., Lv, Z., Gong, M., Jin, Y., Nandi, A.K.: Ultralightweight spatial-spectral feature cooperation network for change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)
45. Li, K., Jiang, J., Han, C., Deng, Y., Chen, K., Zheng, Z., Chen, H., Liu, Z., Gu, Y., Zou, Z., et al.: Open-cd: A comprehensive toolbox for change detection. In: Proceedings of the ACM International Conference on Multimedia. pp. 13608–13612 (2025)
46. Li, Q., Zhong, R., Du, X., Du, Y.: Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2022)
47. Li, Z., Zhou, W., Song, H., Zhang, X., Zhang, K.: Details-enhanced multi-scale network for building change detection in remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **19**, 4530–4544 (2026)
48. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
49. 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)
50. Liu, X., Liu, Y., Jiao, L., Li, L., Liu, F., Yang, S., Hou, B.: Mutsimnet: Mutually reinforcing similarity learning for rs image change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024)
51. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in Neural Information Processing Systems **37**, 103031–103063 (2024)
52. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019 (2022)
53. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
54. Meng, J., Xu, X., Zhang, Z., Li, P., Xie, G., Ren, J., Zheng, Y.: Changeda: Depth-augmented multi-task network for remote sensing change detection via differential analysis. IEEE Transactions on Geoscience and Remote Sensing (2025)
55. Noman, M., Fiaz, M., Cholakkal, H., Narayan, S., Anwer, R.M., Khan, S., Khan, F.S.: Remote sensing change detection with transformers trained from scratch. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
56. Pan, X., Lai, J., Jin, Y., Zhou, X., Zheng, J.: Stenet: A spatial selection and temporal evolution network for change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024)
57. Paranjape, J.N., De Melo, C., Patel, V.M.: A mamba-based siamese network for remote sensing change detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1186–1196. IEEE (2025)
58. Pei, G., Chen, T., Jiang, X., Liu, H., Sun, Z., Yao, Y.: Videomac: Video masked autoencoders meet convnets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22733–22743 (2024)
59. Pei, G., Chen, T., Wang, Y., Cai, X., Shu, X., Zhou, T., Yao, Y.: Seeing what matters: Empowering clip with patch generation-to-selection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24862–24872 (2025)
60. Pei, G., Zhang, L.: Feature hierarchical differentiation for remote sensing image change detection. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022)
61. Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Derczynski, L., Du, X., Grella, M., Gv, K., He, X., Hou, H., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., Lin, J., Mantri, K.S.I., Mom, F., Saito, A., Song, G., Tang, X., Wind, J., Woźniak, S., Zhang, Z., Zhou, Q., Zhu, J., Zhu, R.J.: Rwkv: Reinventing rnns for the transformer era. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 14048–14077 (2023)
62. Peng, B., Goldstein, D., Anthony, Q.G., Albalak, A., Alcaide, E., Biderman, S., Cheah, E., Ferdinan, T., GV, K.K., Hou, H., Krishna, S., Jr., R.M., Muennighoff, N., Obeid, F., Saito, A., Song, G., Tu, H., Zhang, R., Zhao, B., Zhao, Q., Zhu, J., Zhu, R.J.: Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. In: First Conference on Language Modeling (2024)
63. Peng, B., Zhang, R., Goldstein, D., Alcaide, E., Du, X., Hou, H., Lin, J., Liu, J., Lu, J., Merrill, W., et al.: Rwkv-7 "goose" with expressive dynamic state evolution. arXiv preprint arXiv:2503.14456 (2025)

64. Qin, Y., Wang, C., Fan, Y., Pan, C.: Sam2-cd: Remote sensing image change detection with sam2. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 24575–24587 (2025)
65. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
66. Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., Jiang, B.: S2looking: A satellite side-looking dataset for building change detection. Remote Sensing **13**(24), 5094 (2021)
67. Sheng, M., Sun, Z., Cai, Z., Chen, T., Zhou, Y., Yao, Y.: Adaptive integration of partial label learning and negative learning for enhanced noisy label learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4820–4828 (2024)
68. Sheng, M., Sun, Z., Chen, T., Pan, J., Yao, Y., Shen, F.: Revisiting learning with noisy labels: Active forgetting and noise suppression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24792–24802 (2026)
69. Sheng, M., Sun, Z., Chen, T., Pang, S., Wang, Y., Yao, Y.: Foster adaptivity and balance in learning with noisy labels. In: European Conference on Computer Vision. pp. 217–235. Springer (2024)
70. Sheng, M., Sun, Z., Pei, G., Chen, T., Luo, H., Yao, Y.: Enhancing robustness in learning with noisy labels: An asymmetric co-training approach. In: Proceedings of the ACM International Conference on Multimedia. pp. 4406–4415 (2024)
71. Sheng, M., Sun, Z., Zhou, T., Shu, X., Pan, J., Yao, Y.: Ca2c: A prior-knowledge-free approach for robust label noise learning via asymmetric co-learning and co-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 901–911 (2025)
72. Shi, N., Chen, K., Zhou, G.: A divided spatial and temporal context network for remote sensing change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 4897–4908 (2022)
73. Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L.: 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**, 1–16 (2021)
74. Song, Z., Wei, X., Kang, X., Li, S., Liu, J.: Toward efficient remote sensing image change detection via cross-temporal context learning. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–10 (2023)
75. Su, Y., Ma, P., Wang, W., Wang, S., Wu, Y., Li, Y., Jing, P.: Amdanet: Augmented multi-scale difference aggregation network for image change detection. IEEE Transactions on Geoscience and Remote Sensing (2025)
76. Sun, Z., Yao, Y., Liu, T., Li, Z., Shen, F., Tang, J.: Jo-snc: Combating noisy labels through fostering self-and neighbor-consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
77. Tang, X., Zhang, T., Ma, J., Zhang, X., Liu, F., Jiao, L.: Wnet: W-shaped hierarchical network for remote-sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)
78. Thomas, H., Tsai, Y.H.H., Barfoot, T.D., Zhang, J.: Kpconvx: Modernizing kernel point convolution with kernel attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5525–5535 (2024)
79. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: Lsnet: See large, focus small. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9718–9729 (2025)

80. Wang, C., Fan, G., Li, J., Gan, M., Chen, C.P.: Mgcr-net: Multimodal graph-conditioned vision-language reconstruction network for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–15 (2026)
81. Wang, F., Ren, S., Zhang, T., Neskovic, P., Bhattad, A., Xie, C., Yuille, A.: Vit-5: Vision transformers for the mid-2020s. arXiv preprint arXiv:2602.08071 (2026)
82. Wang, J., Song, J., Zhang, H., Zhang, Z., Ji, Y., Zhang, W., Zhang, J., Wang, X.: Spmnet: A siamese pyramid mamba network for very-high-resolution remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing (2025)
83. Wang, L., Fang, Y., Li, Z., Wu, C., Xu, M., Shao, M.: Summator–subtractor network: Modeling spatial and channel differences for change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–12 (2024)
84. 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)
85. Wang, W., Ren, H., Wang, X., Li, J.: Structure sensitive and semantic alignment synergistic enhancement network for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–16 (2025)
86. Xing, Y., Jiang, J., Xiang, J., Yan, E., Song, Y., Mo, D.: Lightcdnet: Lightweight change detection network based on vhr images. IEEE Geoscience and Remote Sensing Letters **20**, 1–5 (2023)
87. Xu, J., Luo, C., Chen, X., Wei, S., Luo, Y.: Remote sensing change detection based on multidirectional adaptive feature fusion and perceptual similarity. Remote Sensing **13**, 3053 (2021)
88. Xu, J., Pei, G., Liu, H., Yao, Y.: Gsv2x: Geometry-aware uncertainty modeling and orthogonal fusion for robust roadside perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21409–21419 (2026)
89. Yang, F., Li, M., Shu, W., Qin, A., Song, T., Gao, C., Xia, G.S.: Convformer-cd: Hybrid cnn-transformer with temporal attention for detecting changes in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing (2025)
90. Yang, Z., Pei, G., Chen, T., Yuan, X., Zhang, H., Shu, X., Yao, Y.: Beyond quadratic: Linear-time change detection with rwkv. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11811–11819 (2026)
91. Yang, Z., Pei, G., Chen, T., Zhou, Y., Zhou, T., Yao, Y., Shen, F.: Efficiency follows global-local decoupling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25524–25535 (2026)
92. Yang, Z., Pei, G., Yao, Y., Zhou, T., Ding, L., Shen, F.: ChangeTitans: Toward remote sensing change detection with neural memory. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2025)
93. Yang, Z., Pei, G., Yao, Y., Zhou, T., Ding, L., Shen, F.: Towards remote sensing change detection with neural memory. arXiv preprint arXiv:2602.10491 (2026)
94. Ye, Y., Wang, M., Zhou, L., Lei, G., Fan, J., Qin, Y.: Adjacent-level feature cross-fusion with 3d cnn for remote sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing (2023)
95. Yin, J., Chen, T., Chen, Y., Pei, G., Shu, X., Yao, Y., Shen, F.: Pca-seg: Revisiting cost aggregation for open-vocabulary semantic and part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27633–27643 (2026)

96. Yin, J., Jiang, X., Chen, T., Pei, G., Yao, Y., Shen, F., Shen, H.T.: Depmatch: Boosting semi-supervised semantic segmentation by exploring depth difference knowledge. IEEE Transactions on Image Processing **35**, 3256–3270 (2026)
97. Yu, H., Chen, H., Jiang, Y., Peng, W., Sun, Z., Kaski, S., Zhao, G.: Attentive convolution: Unifying the expressivity of self-attention with convolutional efficiency. arXiv preprint arXiv:2510.20092 (2025)
98. Yu, W., Zhang, X., Das, S., Zhu, X.X., Ghamisi, P.: Maskcd: A remote sensing change detection network based on mask classification. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
99. 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)
100. Zhang, H., Chen, H., Zhou, C., Chen, K., Liu, C., Zou, Z., Shi, Z.: Bifa: Remote sensing image change detection with bitemporal feature alignment. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024)
101. Zhang, H., Chen, K., Liu, C., Chen, H., Zou, Z., Shi, Z.: Cdmamba: Incorporating local clues into mamba for remote sensing image binary change detection. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–16 (2025)
102. Zhang, H., Lin, M., Yang, G., Zhang, L.: Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images. IEEE Transactions on Neural Networks and Learning Systems **34**, 28–42 (2021)
103. Zhang, K., Zhao, X., Zhang, F., Ding, L., Sun, J., Bruzzone, L.: Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)
104. Zhao, S., Chen, H., Zhang, X., Xiao, P., Bai, L., Ouyang, W.: Rs-mamba for large remote sensing image dense prediction. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
105. Zheng, Z., Zhong, Y., Zhao, J., Ma, A., Zhang, L.: Unifying remote sensing change detection via deep probabilistic change models: From principles, models to applications. ISPRS Journal of Photogrammetry and Remote Sensing **215**, 239–255 (2024)
106. Zhou, B., Li, L., Wang, Y., Liu, H., Yao, Y., Wang, W.: Unialign: Scaling multimodal alignment within one unified model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29644–29655 (2025)
107. Zhou, W., Guo, G., Song, H., Zhang, X., Zhang, K.: Learning boundary-aware semantic context network for remote sensing change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **19**, 4177–4187 (2026)
108. Zhou, Z., Hu, K., Fang, Y., Rui, X.: Schanger: Change detection from a semantic change and spatial consistency perspective. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025)
109. Zhu, L., Wang, X., Zhang, W., Lau, R.: Revisiting the integration of convolution and attention for vision backbone. Advances in Neural Information Processing Systems **37**, 42941–42964 (2024)
110. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: International Conference on Machine Learning. pp. 62429–62442 (2024)