Beyond Quadratic: Linear-Time Change Detection with RWKV

Zhenyu Yang[1,2], Gensheng Pei[3], Tao Chen[1,2], Xia Yuan[1], Haofeng Zhang[1], Xiangbo Shu[1], Yazhou Yao[1,2*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2]State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, China
[3]Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Korea
zhenyu_yang@njust.edu.cn, peigsh@skku.edu, {taochen, yuanxia, zhanghf, shuxb, yazhou.yao}@njust.edu.cn

Abstract

Existing paradigms for remote sensing change detection are caught in a trade-off: CNNs excel at efficiency but lack global context, while Transformers capture long-range dependencies at a prohibitive computational cost. This paper introduces **ChangeRWKV**, a new architecture that reconciles this conflict. By building upon the Receptance Weighted Key Value (RWKV) framework, our ChangeRWKV uniquely combines the parallelizable training of Transformers with the *linear-time* inference of RNNs. Our approach core features two key innovations: a *hierarchical* RWKV encoder that builds multi-resolution feature representation, and a novel *Spatial-Temporal* Fusion Module (STFM) engineered to resolve *spatial* misalignments across scales while distilling fine-grained *temporal* discrepancies. ChangeRWKV not only achieves state-of-the-art performance on the LEVIR-CD benchmark, with an **85.46% IoU** and **92.16% F1** score, but does so while drastically reducing parameters and FLOPs compared to previous leading methods. This work demonstrates a new, efficient, and powerful paradigm for operational-scale change detection. Our code and model are publicly available.

Code — https://github.com/ChangeRWKV/ChangeRWKV

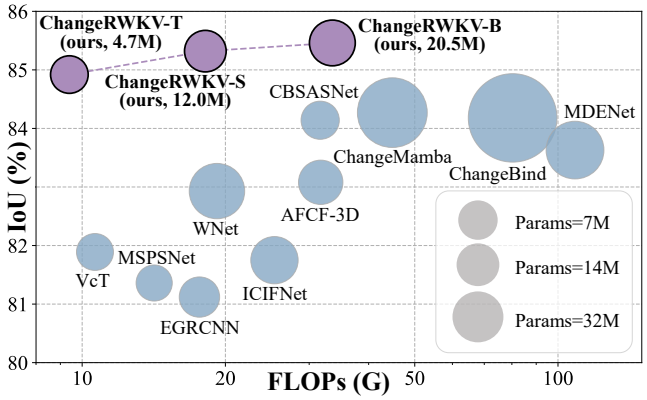

Figure 1: Efficiency versus accuracy on the LEVIR-CD benchmark. The proposed ChangeRWKV family establishes a new state-of-the-art frontier, delivering superior IoU scores while demanding significantly fewer computational resources (FLOPs) and parameters than existing methods. Our tiny model, for instance, achieves a competitive **84.92%** IoU with only **4.7M** Params and **9.40G** FLOPs.

Introduction

Remote sensing change detection (RSCD), the task of identifying meaningful differences from multi-temporal imagery, is fundamental to applications like environmental monitoring, urban planning, and disaster assessment. A fundamental tension defines the recent evolution of RSCD architectures: while high-resolution imagery demands powerful models capable of capturing long-range spatial context, critical applications such as rapid post-disaster analysis on unmanned aerial vehicles (UAVs) require lightweight, low-latency models for on-device deployment.

Early approaches (Zheng et al. 2021; Ye et al. 2023; Huang et al. 2024; Pei et al. 2024; Cai et al. 2024; Pei et al. 2022) based on convolutional neural networks (CNNs) are computationally efficient and excel at extracting local features. However, their inherently local receptive fields limit their ability to model the global context essential for disambiguating complex changes. Vision Transformers (ViTs) (Dosovitskiy et al. 2021; Liu et al. 2021; Xie et al. 2021; Bernhard, Strauß, and Schubert 2023; Chen et al. 2024b; Zhou et al. 2025) overcame this limitation by using self-attention to model global relationships, leading to significant accuracy gains (Chen et al. 2024b; Zang et al. 2025; Benidir, Gonthier, and Mallet 2025). This performance, however, comes at the cost of quadratic complexity, rendering standard ViTs impractical for the high-resolution images common in remote sensing. This bottleneck has spurred research into linear-time architectures, such as state-space models like Mamba (Gu and Dao 2024; Dao and Gu 2024), which promise both global modeling and linear scalability.

Recently, the Receptance Weighted Key Value (RWKV) architecture (Peng et al. 2023, 2024, 2025) has emerged as a compelling alternative, blending the parallelizable training of Transformers with the linear complexity of RNNs. Unlike self-attention with its $\mathcal{O}(T^2d)$ complexity, RWKV operates with $\mathcal{O}(Td)$ complexity in both computation and memory, making it highly scalable. Its design, which decouples spatial and channel mixing, has demonstrated competitive performance and enhanced training stability over other linear-

*Corresponding Author

time models. While RWKV's potential has been realized in natural language and, more recently, in general vision tasks (Duan et al. 2025; Zhou and Chen 2024; Jiang et al. 2025; Lv and Zhang 2025), its aptitude for the nuanced task of bi-temporal image analysis in RSCD remains unexplored.

In this work, we bridge this gap by presenting ChangeRWKV, a robust and efficient framework for remote sensing change detection. Our core idea is to leverage the linear complexity and strong representation power of RWKV to build a powerful yet lightweight model. We design a hierarchical encoder that processes each image to extract rich, multi-scale features with minimal computational overhead. At the heart of our model is a novel Spatial-Temporal Fusion Module (STFM), specifically engineered to effectively integrate these multi-scale features and model the temporal discrepancies between the image pairs. As shown in Fig. 1, ChangeRWKV sets a new standard for the trade-off between accuracy and efficiency, making high-performance change detection feasible even under constrained computational budgets. Our main contributions are summarized as follows:

- We propose ChangeRWKV, the *first* framework to successfully adapt the RWKV architecture for remote sensing change detection, establishing a new benchmark for highly *efficient* yet *accurate* models.
- We introduce a novel STFM that adeptly integrates hierarchical features and models bi-temporal differences, significantly enhancing the model's ability to discriminate subtle and complex changes.
- We validate our approach through extensive experiments on four diverse optical and SAR change detection benchmarks. Our results show that ChangeRWKV achieves state-of-the-art performance while drastically reducing computational costs, confirming its suitability for real-time and resource-limited deployment scenarios.

Related Work

CNN-based Change Detection. CNNs form the foundation of deep learning-based RSCD. Early methods primarily utilize fully convolutional Siamese architectures to extract and compare pixel-wise features from bi-temporal images (Daudt, Le Saux, and Boulch 2018). To improve performance, subsequent works introduce more sophisticated designs, such as nested U-Nets for enhanced multi-scale feature preservation (Fang et al. 2021), dual-decoders to separate change and semantic information (Chen et al. 2022), and specialized modules for edge-aware fusion or 3D spatio-temporal modeling (Huang et al. 2024; Ye et al. 2023). While computationally efficient, the performance of CNN-based models is inherently constrained by their local receptive fields, which struggle to capture the global context necessary for understanding complex semantic changes. Our approach overcomes this by employing a backbone capable of long-range dependency modeling.

Transformer-based Change Detection. To address the context limitations of CNNs, ViTs are adapted for RSCD. Seminal works like BIT (Chen, Qi, and Shi 2021) and ChangeFormer (Bandara and Patel 2022) demonstrate the power of self-attention in modeling global spatial-temporal relationships, leading to significant accuracy improvements. This paradigm is further refined through hierarchical designs like SwinSUNet (Zhang et al. 2022) and hybrid models that combine convolutional priors with attention mechanisms (Pei and Zhang 2022; Yang et al. 2025). Although Transformers excel at global context modeling, their quadratic complexity in self-attention creates a severe computational bottleneck, limiting their practical use on high-resolution imagery and in resource-constrained settings (Mao et al. 2025). Our work directly targets this efficiency gap without sacrificing the ability to model long-range interactions.

Linear-Time Change Detection. Most recently, the field has gravitated towards linear-time architectures to balance modeling power and efficiency. This trend is dominated by methods adapting State Space Models (SSMs), particularly Mamba (Gu and Dao 2024), for the change detection task. Architectures like ChangeMamba (Chen et al. 2024a), CD-Mamba (Zhang et al. 2025a), and RS-Mamba (Zhao et al. 2024) successfully integrate Mamba's selective scan mechanism for efficient sequential modeling of image patches. Other variants explore frequency-domain analysis (Xing et al. 2025) or change-guided feature selection (Ghazaei and Aptoula 2025) within the SSM framework. These Mamba-based models confirm the viability of linear-time architectures for RSCD, yet the exploration of this design space remains narrow. Our work diverges by being the first to investigate the RWKV architecture (Peng et al. 2023), whose linear complexity, training stability, and architectural simplicity present a novel and powerful alternative for highly efficient remote sensing change detection.

Method

The RWKV Architecture

The Receptance Weighted Key Value (RWKV) model (Peng et al. 2023, 2024, 2025) is a novel sequence architecture that marries the parallelizable training of Transformers with the linear inference cost of RNNs. It reformulates the self-attention mechanism, which has a complexity of $\mathcal{O}(T^2d)$, into a recurrent structure with a linear complexity of $\mathcal{O}(Td)$ *w.r.t.* sequence length T, and model dimension d.

At its core, RWKV processes a sequence of tokens $\{x_t\}_{t=1}^{T}$. For each token x_t, it first projects it into receptance (r_t), key (k_t), and value (v_t) vectors. This projection cleverly incorporates information from the previous token x_{t-1} through a learnable interpolation:

$$(*)_t = W_{(*)} \cdot \left[\mu_{(*)} \odot x_t + (1 - \mu_{(*)}) \odot x_{t-1}\right], \quad (1)$$

where $(*) \in \{r, k, v\}$, and $\odot$ denotes Hadamard product. $W_{(*)} \in \mathbb{R}^{d\times d}$ are the projection matrices, and $\mu_{(*)} \in \mathbb{R}^d$ are the learnable mixing coefficients. The central mechanism is the `WKV` operation, which acts as a time-decaying, channel-wise aggregation of past values, modulated by keys:

$$\texttt{WKV}_t = \frac{\sum_{i=1}^{t-1} e^{-(t-1-i)w+k_i} \odot v_i + e^{u+k_t} \odot v_t}{\sum_{i=1}^{t-1} e^{-(t-1-i)w+k_i} + e^{u+k_t}}, \quad (2)$$

where $w \in \mathbb{R}^d$ is a learnable decay vector that controls how much past information is forgotten, and $u \in \mathbb{R}^d$ is a learnable parameter that grounds the current token's importance. The RWKV block consists of two main sub-layers:

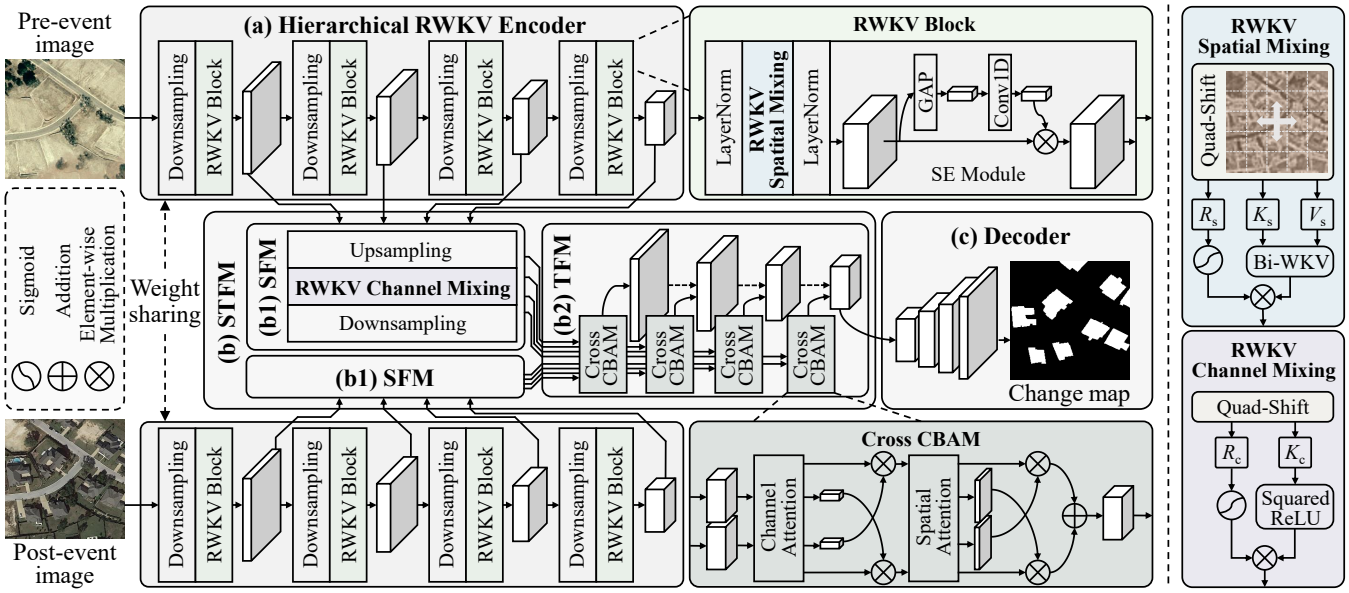

Figure 2: Overall architecture of ChangeRWKV. The model consists of three main components: **(a)** a Hierarchical RWKV Encoder that extracts multi-scale features, **(b)** a Spatial-Temporal Fusion Module (STFM) that integrates spatial and temporal cues, and **(c)** a lightweight Decoder that generates the change mask. The STFM is further decomposed into **(b1)** a Spatial Fusion Module (SFM) for multi-scale spatial alignment and **(b2)** a Temporal Fusion Module (TFM) for bi-temporal interaction.

Time-Mixing models sequential dependencies. The output is gated by the receptance r_t, allowing the model to dynamically control how much historical context to incorporate:

$$o_t = W_o \cdot (\sigma(r_t) \odot \texttt{WKV}_t), \tag{3}$$

where $W_o \in \mathbb{R}^{d \times d}$ is an output projection matrix, and $\sigma(\cdot)$ is the `Sigmoid` function for modulating historical context.
Channel-Mixing is applied per-token to model feature interactions within each token's embedding. It is typically implemented as a simple two-layer `MLP` with a squared `ReLU` activation for enhanced non-linearity. This decoupled design of time and channel mixing contributes to RWKV's training stability and strong performance.

Overall Architecture of ChangeRWKV

Given a pair of co-registered remote sensing images $I_A, I_B \in \mathbb{R}^{H \times W \times C}$ captured at times t_A and t_B, the objective of RSCD is to predict a binary change map $M \in \{0,1\}^{H \times W}$, where $M_{m,n} = 1$ if a change occurred at pixel (m, n) and 0 otherwise. As shown in Fig. 2, our ChangeRWKV follows a Siamese encoder-decoder structure.
Hierarchical RWKV Encoder: A shared RWKV-based vision encoder processes I_A and I_B independently to extract hierarchical feature maps $\{f_{i1}, f_{i2}, f_{i3}, f_{i4}\}$ for each image $i \in \{A, B\}$ at four different scales.
Spatial-Temporal Fusion Module: In this work, we introduce a novel STFM, which processes the input bi-temporal, multi-scale features by first performing intra-image spatial fusion and then inter-image temporal fusion, yielding a set of change-centric feature maps $\{\tilde{f}_1, \tilde{f}_2, \tilde{f}_3, \tilde{f}_4\}$.
Lightweight Decoder: A U-Net (Ronneberger, Fischer, and Brox 2015) style decoder with skip connections takes the fused features $\{\tilde{f}_1, \tilde{f}_2, \tilde{f}_3, \tilde{f}_4\}$ and progressively upsamples them to generate the final change map $\hat{M}$.

Hierarchical RWKV Encoder

We adapt the sequential RWKV block for 2D vision tasks. Following the principles of VisionRWKV (Duan et al. 2025), we replace the unidirectional time-mixing with a bidirectional spatial-mixing mechanism that aggregates information across the 2D plane, enhancing local context modeling. For efficiency, the standard channel-mixing `MLP` is replaced by a lightweight Squeeze-and-Excitation (SE) module (Hu, Shen, and Sun 2018). Crucially, our encoder produces *hierarchical* outputs, providing the *multi-scale* representations essential for detecting changes of varying sizes, a key departure from flat-feature vision models.

Spatial-Temporal Fusion Module

The STFM is the core of our method, designed to robustly integrate features across both space and time (see Fig. 2(b)).
Spatial Fusion Module (SFM): This module enriches features within each temporal snapshot by promoting cross-scale communication. For each image $i \in \{A, B\}$, its feature maps $\{f_{ij}\}_{j=1}^{4}$ are all upsampled to the finest resolution level ($H/2 \times W/2$) and concatenated along the channel dimension into a single tensor F_i. A residual channel-mixing block then refines this tensor:

$$\hat{F}_i = F_i + \texttt{Channel-Mixing}(F_i). \tag{4}$$

The refined tensor $\hat{F}_i$ is subsequently split and downsampled back to the original four scales, yielding spatially-enhanced features $\{\tilde{f}_{ij}\}_{j=1}^{4}$ for each image $i \in \{A, B\}$.

Method	Params (M)	FLOPs (G)	LEVIR-CD				WHU-CD			
			IoU	F1	P	R	IoU	F1	P	R
FC-Siam-Diff (2018)	4.38	1.35	75.92	86.31	89.53	83.31	41.66	58.81	47.33	77.66
FC-Siam-Conc (2018)	4.99	1.55	71.96	83.69	91.99	76.77	49.95	66.63	60.88	73.58
SNUNet (2021)	12.03	27.44	78.83	88.16	89.18	87.17	71.67	83.50	85.60	81.49
BIT (2021)	3.55	4.35	80.68	89.31	89.24	89.37	72.39	83.98	86.64	81.48
ICIFNet (2022)	23.82	25.36	81.75	89.96	91.32	88.64	79.24	88.32	92.98	85.56
ChangeFormer (2022)	41.02	202.80	82.48	90.40	92.05	88.80	71.91	83.66	85.49	81.90
WNet (2023)	43.07	19.20	82.93	90.67	91.16	90.18	83.91	91.25	92.37	90.15
AFCF-3D (2023)	17.54	31.72	83.08	90.76	91.35	90.17	87.93	93.58	93.47	93.69
CCLNet++ (2023)	28.78	23.27	83.62	91.08	92.31	89.88	82.32	90.31	89.83	90.78
CF-GCN (2024)	13.58	43.93	83.41	90.96	91.75	90.18	84.90	91.83	94.81	89.04
SEIFNet (2024)	8.37	27.91	83.40	90.95	92.49	89.46	76.04	86.39	87.01	85.77
BiFA (2024)	5.58	53.00	82.96	90.69	91.52	89.86	89.34	94.37	95.15	<u>93.60</u>
CBSASNet (2024)	5.76	31.64	84.14	91.39	92.47	90.33	86.08	92.52	93.93	91.15
ChangeBind (2024)	153.70	80.29	84.18	91.41	**94.93**	88.14	85.72	92.31	94.81	89.94
ChangeMamba (2024a)	84.70	44.86	84.27	91.37	<u>93.87</u>	89.00	88.02	93.63	94.22	93.05
ConvFormer-CD/48 (2025)	37.72	5.14	84.23	91.44	92.37	90.52	85.41	92.13	95.14	89.26
SPMNet (2025)	17.33	6.94	84.07	90.99	92.12	90.58	84.84	91.80	94.67	89.10
SFEARNet (2025)	5.56	4.65	83.23	90.85	91.43	90.27	85.81	92.36	94.25	90.55
STRobustNet (2025b)	5.23	13.19	83.66	91.11	91.54	90.67	83.29	90.89	92.92	88.94
AMDANet (2025)	25.83	77.23	82.34	90.32	92.45	88.28	80.68	89.30	88.77	89.84
ChangeRWKV-T (ours)	4.66	9.40	84.92	91.85	92.41	<u>91.29</u>	88.19	93.72	**96.27**	91.30
ChangeRWKV-S (ours)	12.00	18.15	<u>85.32</u>	<u>92.08</u>	93.00	91.17	**90.06**	**94.77**	95.50	**94.05**
ChangeRWKV-B (ours)	20.50	33.56	**85.46**	**92.16**	92.86	**91.46**	<u>89.59</u>	<u>94.51</u>	<u>96.14</u>	92.93

Table 1: Quantitative comparison on the LEVIR-CD (Chen and Shi 2020) and WHU-CD (Hong et al. 2023) test sets. All metrics are reported as a percentage (%), with the highest value highlighted in **bold**, and the second highest <u>underlined</u>.

Temporal Fusion Module (TFM): After spatial enhancement, the TFM integrates the bi-temporal features at each scale j. Inspired by Convolutional Block Attention Module (CBAM) (Woo et al. 2018), we employ a new cross-attention strategy, termed as Cross CBAM (see Fig. 2), to explicitly model changes. First, channel attention weights are computed from one temporal feature and applied to the other, and vice-versa, to highlight discriminative channels.

$$\gamma_{ij}^{c} = \sigma(\texttt{MLP}(\texttt{GAP}(\tilde{f}_{ij}))), \tag{5}$$

where GAP denotes Global Average Pooling. The features are then cross-refined: $\tilde{f}'_{Aj} = \gamma^{c}_{Bj} \odot \tilde{f}_{Aj}$ and $\tilde{f}'_{Bj} = \gamma^{c}_{Aj} \odot \tilde{f}_{Bj}$. Next, spatial attention maps are computed from these channel-refined features and are also cross-applied to focus on salient spatial regions of change:

$$s_{ij} = \sigma(\texttt{Conv}([\texttt{AvgPool}(\tilde{f}'_{ij}); \texttt{MaxPool}(\tilde{f}'_{ij})])). \tag{6}$$

The final fused feature at each level j is obtained by element-wise summation of the fully refined temporal features: $\tilde{f}_j = (\tilde{f}'_{Aj} \odot s_{Bj}) + (\tilde{f}'_{Bj} \odot s_{Aj})$. This adaptive, data-driven fusion contrasts with methods relying on simple subtraction (Daudt, Le Saux, and Boulch 2018; Corley, Robinson, and Ortiz 2024) or predefined metrics (Wang et al. 2022; Dong et al. 2024), allowing our model to learn optimal fusion strategies for diverse types of change.

Loss Function

To effectively train ChangeRWKV, we employ a hybrid loss function that combines Binary Cross Entropy (BCE) and Dice loss. The BCE loss ensures pixel-level accuracy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{n=1}^{N}[M_n \log(\hat{M}_n) + (1 - M_n)\log(1 - \hat{M}_n)], \tag{7}$$

where $N = H \times W$, M is the ground-truth map, and $\hat{M}$ is the prediction. To address class imbalance and improve the segmentation of change region boundaries, we add the Dice loss, which maximizes the spatial overlap:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum M_n \hat{M}_n + \epsilon}{\sum M_n + \sum \hat{M}_n + \epsilon}, \tag{8}$$

where ϵ is a small constant for numerical stability. The final training objective is a weighted sum of these two losses:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{Dice}}, \tag{9}$$

where λ is a hyperparameter balancing the two terms, this composite loss promotes both high pixel-wise fidelity and structurally coherent change maps.

Experiments

Experimental Setup

Datasets. We conduct extensive evaluations on four public benchmarks spanning both optical and synthetic aperture radar (SAR) modalities to demonstrate the effectiveness and generalization of our method. Details are as follows:

- **LEVIR-CD** (Chen and Shi 2020): A widely-used dataset of 637 high-resolution (0.5m/pixel) image pairs focused on building changes in urban environments.

Method	Params (M)	FLOPs (G)	IoU	F1
FC-Siam-Diff (2018)	4.38	1.35	57.44	72.97
FC-Siam-Conc (2018)	4.99	1.55	58.07	73.48
STANet (2020)	16.93	6.58	65.66	79.31
SNUNet (2021)	12.03	27.44	67.11	80.32
BIT (2021)	3.55	4.35	70.64	82.80
TransUNetCD (2022)	28.37	244.54	71.86	83.63
SwinSUNet (2022)	39.28	43.50	74.82	85.60
CTDFormer (2023)	3.85	303.77	67.09	80.30
BiFA (2024)	5.58	53.00	72.35	83.96
ConvFormer-CD/48 (2025)	37.72	5.14	74.37	85.30
SPMNet (2025)	17.33	6.94	71.01	83.09
ChangeRWKV-T (ours)	4.66	9.40	75.14	85.80
ChangeRWKV-S (ours)	12.00	18.15	75.21	85.85
ChangeRWKV-B (ours)	20.50	33.56	**75.46**	**86.01**

Table 2: Quantitative comparison on the LEVIR-CD+ (Shen et al. 2021) test set. All metrics are reported as a percentage (%). Best results are highlighted in **bold**.

Method	Params (M)	FLOPs (G)	IoU	F1
FC-Siam-Diff (2018)	4.38	1.35	86.07	92.52
FC-Siam-Conc (2018)	4.99	1.55	93.39	96.58
BIT (2021)	3.55	4.35	93.48	96.63
MFPNet (2021)	60.06	107.73	97.09	98.52
MSCANet (2022)	16.59	14.68	94.43	97.14
FCS (2022)	3.03	9.59	94.77	97.31
DED (2022)	3.10	9.90	94.94	97.40
FCCDN (2022)	6.31	12.52	95.15	97.51
ChangeFormer (2022)	41.02	202.80	96.47	98.20
AFCF-3D (2023)	17.54	31.72	93.72	96.76
ChangeMamba (2024a)	84.70	44.86	90.40	94.96
ChangeRWKV-T (ours)	4.66	9.40	96.87	98.41
ChangeRWKV-S (ours)	12.00	18.15	97.02	98.49
ChangeRWKV-B (ours)	20.50	33.56	**97.18**	**98.57**

Table 3: Quantitative comparison on the SAR-CD (Alatalo, Sipola, and Rantonen 2023) test set. All metrics are reported as a percentage (%). Best results are highlighted in **bold**.

- **WHU-CD** (Hong et al. 2023): Comprises two large aerial images (0.3m/pixel) capturing significant structural evolution in Christchurch, New Zealand.
- **LEVIR-CD+** (Shen et al. 2021): An extension of LEVIR with 985 image pairs featuring longer time spans (5-14 years), and more diverse, complex change patterns.
- **SAR-CD** (Alatalo, Sipola, and Rantonen 2023): A challenging dataset of 10,000 512×512 Sentinel-1 SAR image pairs with simulated changes, testing model robustness against speckle noise and imaging geometries.

Implementation Details. Following the setup of (Li et al. 2024), all image pairs are cropped into fixed-size patches of 256×256 using a sliding window strategy. Our models are optimized with the Adam optimizer (Kingma and Ba 2015), using an initial learning rate of 10^{-5}, a weight decay of 10^{-4}, and a batch size of 8 per GPU. All experiments are conducted on four NVIDIA RTX 3090 GPUs. The learning rate follows a cosine annealing schedule, preceded by a warm-up phase of 20 epochs. Training proceeds for 200 epochs with a gradient clipping threshold of 0.5 to ensure numerical stability. To improve generalization, we adopt standard data augmentation techniques, including random rotation, flipping, brightness-contrast adjustment, and Gaussian blur, following FHD (Pei and Zhang 2022).

To explore the trade-off between accuracy and efficiency, we instantiate ChangeRWKV in three model scales: ChangeRWKV-T (**T**iny), -S (**S**mall), and -B (**B**ase), which differ in embedding dimension and encoder depth. Specifically, their configurations are as follows:

- **ChangeRWKV-T** uses embedding dimensions of [32, 48, 96, 160] and encoder depths of [2, 2, 4, 2].
- **ChangeRWKV-S** uses embedding dimensions of [32, 64, 128, 192] and encoder depths of [3, 3, 6, 3].
- **ChangeRWKV-B** uses embedding dimensions of [48, 72, 144, 240] and encoder depths of [3, 3, 6, 3].

Evaluation Metrics. We use four standard metrics for quantitative evaluation: Precision (P), Recall (R), F1-score (F1), and Intersection over Union (IoU). IoU and F1 are the primary indicators of overall performance.

Comparison with State-of-the-Art Methods

Benchmarked against leading state-of-the-art (SOTA) methods across four diverse datasets, ChangeRWKV sets a new standard in accuracy while establishing a superior trade-off between performance and computational cost.

Performance on Standard Optical Benchmarks. As shown in Table 1, our ChangeRWKV-B sets a new SOTA on the widely-used LEVIR-CD dataset, with an IoU of 85.46% and an F1 of 92.16%. This surpasses strong recent methods like ChangeBind (Noman, Fiaz, and Cholakkal 2024) (84.18% IoU) and CBSASNet (He et al. 2024) (84.14% IoU). More importantly, our lightweight ChangeRWKV-T model, with just **4.66M** parameters and **9.40G** FLOPs, already outperforms most prior work with an IoU of **84.92%**. On the WHU-CD dataset, our models maintain this strong performance, with the mid-sized ChangeRWKV-S achieving a top IoU of **90.06%** and an F1 of **94.77%**. These results confirm the effectiveness and scalability of the ChangeRWKV architecture on high-resolution optical imagery.

Robustness to Long-Term Temporal Variations. The LEVIR-CD+ dataset presents a more challenging test with longer time intervals between images. As shown in Table 2, ChangeRWKV excels in this setting. ChangeRWKV-B leads with a **75.46%** IoU and an **86.01%** F1, outperforming much heavier Transformer models like SwinSUNet (39.28M params) and ConvFormer-CD (37.72M params) while using only 20.5M parameters. The ability of our models to handle diverse, long-term changes underscores the effectiveness of our spatial-temporal fusion design and the representational power of the linear-attention backbone.

Generalization to SAR Imagery. As detailed in Table 3, we evaluate the fundamental difference-learning capability of our models on SAR-CD, which features synthetic changes and significant speckle noise. Despite being designed for optical data, ChangeRWKV demonstrates remarkable general-

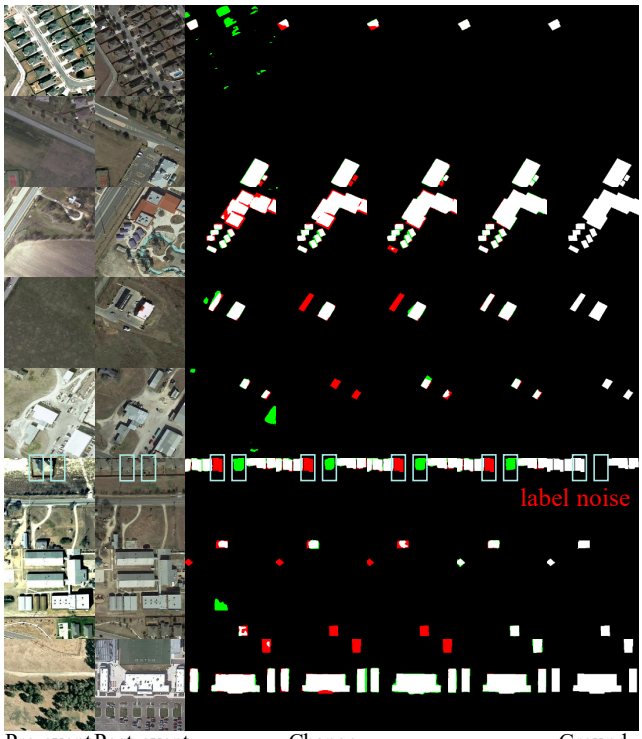

Figure 3: Qualitative results on the LEVIR-CD test set. Predicted outputs are color-coded as follows: white for true positives (TP), black for true negatives (TN), green for false positives (FP), and red for false negatives (FN).

ization. The base model sets a new benchmark with an IoU of 97.18% and an F1 of 98.57%. Even the Tiny and Small variants surpass previous SOTA methods with a fraction of the parameters. This strong performance, without specific tuning for SAR data, highlights that our model learns a robust and modality-agnostic representation of change.

Qualitative Analysis. As shown in Fig. 3, visual comparisons on LEVIR-CD confirm the superiority of our method. ChangeRWKV generates markedly sharper and more complete change masks, particularly in challenging scenarios where other methods struggle. It precisely delineates the boundaries of small, irregular buildings in dense urban layouts and maintains high fidelity for objects near image borders, a testament to the robustness of our hierarchical encoder and fusion module. Furthermore, our model effectively suppresses spurious predictions arising from background clutter and demonstrates resilience to common annotation noise (Sheng et al. 2024a,b), resulting in cleaner and more reliable change maps. This robustness extends to different sensing modalities. Fig. 4 illustrates this on SAR-CD, where ChangeRWKV produces clean and accurate detections despite significant speckle noise and non-semantic changes. This ability to perform well under heterogeneous conditions highlights the strong generalization of our framework, which learns fundamental patterns of change independent of the sensor type.

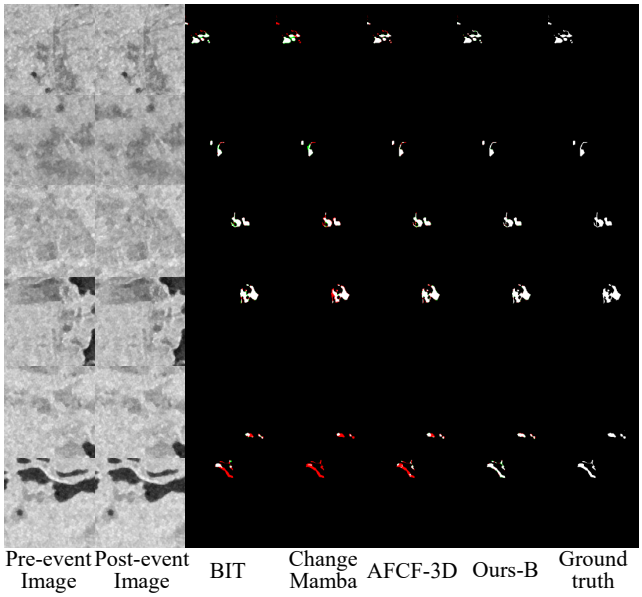

Figure 4: Qualitative results on the SAR-CD test set. Predicted outputs are color-coded as follows: white for true positives (TP), black for true negatives (TN), green for false positives (FP), and red for false negatives (FN).

Ablation and Analysis

We conduct a series of detailed experiments on LEVIR-CD to dissect the ChangeRWKV architecture and validate our design choices. We analyze the contribution of each core component and rigorously evaluate the model's computational efficiency, with results summarized in Table 4.

Encoder Architecture (*cf.* ❶ ❷). To validate our hierarchical design, we compare it against two baselines: a "flat" VisionRWKV encoder (Duan et al. 2025) and the original VisionRWKV design which uses a ViT-Adapter (Chen et al. 2023) for segmentation. The results show a clear progression: simply adding the adapter for hierarchical features improves the IoU from 81.86% to **83.06%**. Our tailored hierarchical encoder further boosts performance to **85.46%** IoU and **92.16%** F1, confirming that a specialized multi-scale architecture is essential for effective change detection.

Channel Mixing Module (*cf.* ❸ ❹). We analyze the trade-off between performance and computational cost in our channel mixing design. Compared to our lightweight SE-based module, a heavier, standard RWKV-style mixing block provides a negligible performance difference (85.38% vs. **85.46%** IoU). However, our design is vastly more efficient, reducing computational cost from 71.66G to **33.56G** FLOPs and parameters from 47.1M to **20.5M**. This demonstrates that our lightweight module effectively captures feature interactions at less than half the computational budget.

Spatial Fusion Module (*cf.* ❺ ❻). The SFM is designed to align features across scales before temporal comparison. Removing SFM or replacing it with an FPN-style (Lin et al. 2017) fusion degrades IoU by **2.82%** and **1.90%**, respectively, highlighting that enforcing intra-image spatial consistency is critical for precise change localization.

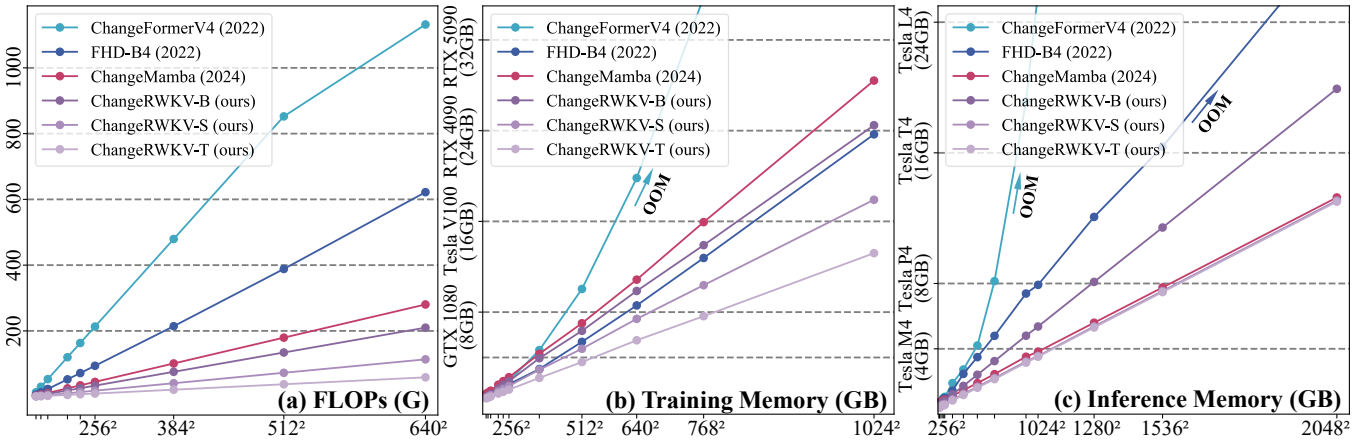

Figure 5: Computational Scalability Analysis. ChangeRWKV demonstrates near-linear growth in (a) FLOPs, (b) training memory, and (c) inference memory, significantly outperforming Transformer-based models at high resolutions.

Experiment	Variant #	Method	IoU	F1
Component				
Encoder Architectures	❶	VisionRWKV	81.86	90.03
	❷	+ ViT-Adapter	83.06	90.75
	-	**ours**	**85.46**	**92.16**
Channel Mixing Module	❸	*w/o*	84.43	91.56
	❹	RWKV-like	85.38	92.11
	-	**ours**	**85.46**	**92.16**
Spatial Fusion Module	❺	*w/o*	82.64	90.49
	❻	FPN	83.56	91.05
	-	**ours**	**85.46**	**92.16**
Temporal Fusion Module	❼	SiamDiff	82.68	90.52
	❽	SiamConc	82.73	90.55
	❾	FHD	85.01	91.90
	❿	STRM	83.13	90.79
	-	**ours**	**85.46**	**92.16**
Hyperparameter				
Balance Factor λ	①	0	84.95	91.86
	②	0.5	85.38	92.11
	-	**1.0 (ours)**	**85.46**	**92.16**
	③	2.0	85.12	91.96

Table 4: Ablation study on different architectural components. All metrics are reported as a percentage (%).

Temporal Fusion Module (*cf.* ❼ ❽ ❾ ❿). We benchmark our cross-attention based TFM against several common fusion strategies. Our module significantly outperforms simpler methods like SiamDiff (82.68% IoU) and SiamConc (82.73% IoU), as well as more recent techniques like FHD (Pei and Zhang 2022) (85.01% IoU) and STRM (Chen et al. 2024a) (83.13% IoU). The superior performance of our TFM highlights its advanced capability to model complex, fine-grained temporal interactions between feature pairs.

Loss Function Balance (*cf.* ① ② ③). We finally examine the weighting factor λ for the Dice loss component. We found that a balanced contribution ($\lambda = 1.0$) provides the best results, confirming that a joint objective optimizing both pixel-level accuracy (BCE) and region-level coherence (Dice) is optimal for the remote sensing change detection task.

Efficiency and Scalability Analysis

A core motivation for ChangeRWKV is to address the computational bottleneck in high-resolution change detection. We evaluate our models' resource consumption for inputs ranging from 64^2 to 2048^2 pixels. As shown in Fig. 5, ChangeRWKV's linear-time attention mechanism provides a clear advantage. While Transformer-based methods exhibit quadratic growth, our models' resource usage scales near-linearly. This efficiency is stark in direct comparison: at a 512×512 resolution, our ChangeRWKV-B requires only **134.25G** FLOPs and **6.5GB** of training memory, a sharp contrast to the 852.44G FLOPs and 10.2GB required by ChangeFormer. This makes our approach vastly more scalable. Critically, this efficiency translates to real-world deployability. All three ChangeRWKV variants can perform inference on 1024^2 inputs using a single NVIDIA Tesla P4 GPU with only 8GB of VRAM, highlighting its suitability for resource-constrained and edge-computing scenarios.

Conclusion

This paper presents ChangeRWKV, an *efficient* and *powerful* framework for remote sensing change detection that directly addresses the trade-off between accuracy and computational cost. Our approach leverages a hierarchical encoder built upon the *linear-time* RWKV architecture to capture multi-scale spatial features with exceptional efficiency. The core of our framework is the novel Spatial-Temporal Fusion Module (STFM), which adaptively integrates bi-temporal features to highlight meaningful changes. Comprehensive experiments show that ChangeRWKV establishes a new state-of-the-art on multiple optical and SAR benchmarks, delivering superior performance while requiring only a fraction of the computational resources of previous methods.

Limitations and Future Work. Despite its strong performance, our work's limitations include the reliance on large annotated datasets and a lack of modality-specific priors (*e.g.*, for SAR or hyperspectral data). Future work will focus on exploring weakly-supervised learning and enabling real-time deployment on resource-constrained platforms.

Acknowledgments

This work was supported by the National Defense Science and Technology Industry Bureau Technology Infrastructure Project (JSZL2024606C001), Key Research and Development (R&D) Plan Project of Jiangsu Province (No. BE2023008-2).

References

Alatalo, J.; Sipola, T.; and Rantonen, M. 2023. Improved Difference Images for Change Detection Classifiers in SAR Imagery Using Deep Learning. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–14.

Bandara, W. G. C.; and Patel, V. M. 2022. A Transformer-Based Siamese Network for Change Detection. In *IGARSS*, 207–210. IEEE.

Benidir, Y.; Gonthier, N.; and Mallet, C. 2025. The Change You Want To Detect: Semantic Change Detection In Earth Observation With Hybrid Data Generationf. In *CVPR*, 2204–2214.

Bernhard, M.; Strauß, N.; and Schubert, M. 2023. MapFormer: Boosting Change Detection by Using Pre-change Information. In *ICCV*, 16837–16846.

Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; and Yao, Y. 2024. Poly kernel inception network for remote sensing detection. In *CVPR*, 27706–27716.

Chen, H.; Qi, Z.; and Shi, Z. 2021. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–14.

Chen, H.; and Shi, Z. 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.*, 12(10): 1662.

Chen, H.; Song, J.; Han, C.; Xia, J.; and Yokoya, N. 2024a. ChangeMamba: Remote Sensing Change Detection with Spatiotemporal State Space Model. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–20.

Chen, H.; Xu, T.; Chen, Z.; Liu, P.; Bai, H.; and Li, J. 2024b. Multi-Scale Change-Aware Transformer for Remote Sensing Image Change Detection. In *ACM MM*, 2992–3000.

Chen, P.; Zhang, B.; Hong, D.; Chen, Z.; Yang, X.; and Li, B. 2022. FCCDN: Feature Constraint Network for VHR Image Change Detection. *ISPRS J. Photogramm. Remote Sens.*, 187: 101–119.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2023. Vision Transformer Adapter for Dense Predictions. In *ICLR*.

Corley, I.; Robinson, C.; and Ortiz, A. 2024. A Change Detection Reality Check. *arXiv preprint arXiv:2402.06994*.

Dao, T.; and Gu, A. 2024. Transformers Are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *ICML*. PMLR.

Daudt, R. C.; Le Saux, B.; and Boulch, A. 2018. Fully Convolutional Siamese Networks for Change Detection. In *ICIP*, 4063–4067. IEEE.

Dong, S.; Wang, L.; Du, B.; and Meng, X. 2024. ChangeCLIP: Remote Sensing Change Detection with Multimodal Vision-Language Representation Learning. *ISPRS J. Photogramm. Remote Sens.*, 208: 53–69.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Duan, Y.; Wang, W.; Chen, Z.; Zhu, X.; Lu, L.; Lu, T.; Qiao, Y.; Li, H.; Dai, J.; and Wang, W. 2025. Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures. In *ICLR*.

Fang, S.; Li, K.; Shao, J.; and Li, Z. 2021. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geosci. Remote Sens. Lett.*, 19: 1–5.

Feng, Y.; Xu, H.; Jiang, J.; Liu, H.; and Zheng, J. 2022. ICIF-Net: Intra-Scale Cross-Interaction and Inter-Scale Feature Fusion Network for Bitemporal Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–13.

Ghazaei, E.; and Aptoula, E. 2025. Change State Space Models for Remote Sensing Change Detection. *arXiv preprint arXiv:2504.11080*.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *CoLM*.

He, N.; Wang, L.; Zheng, P.; Zhang, C.; and Li, L. 2024. CBSAS-Net: A Siamese Network Based on Channel Bias Split Attention for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–17.

Hong, D.; Qiu, C.; Yu, A.; Quan, Y.; Liu, B.; and Chen, X. 2023. Multi-Task Learning for Building Extraction and Change Detection from Remote Sensing Images. *Appl. Sci.*, 13(2): 1037.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *CVPR*, 7132–7141.

Huang, Y.; Li, X.; Du, Z.; and Shen, H. 2024. Spatiotemporal Enhancement and Interlevel Fusion Network for Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–14.

Jiang, J.; Zhang, J.; Liu, W.; Gao, M.; Hu, X.; Yan, X.; Huang, F.; and Liu, Y. 2025. RWKV-UNet: Improving UNet with Long-Range Cooperation for Effective Medical Image Segmentation. *arXiv preprint arXiv:2501.08458*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Li, K.; Jiang, J.; Codegoni, A.; Han, C.; Deng, Y.; Chen, K.; Zheng, Z.; Chen, H.; Zou, Z.; Shi, Z.; Fang, S.; Meng, D.; Wang, Z.; and Cao, X. 2024. Open-CD: A Comprehensive Toolbox for Change Detection. *arXiv preprint arXiv:2407.15317*.

Li, M.; Ming, D.; Xu, L.; Dong, D.; and Zhang, Y. 2025. SFEAR-Net: A Network Combining Semantic Flow and Edge-Aware Refinement for Highly Efficient Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–18.

Li, Q.; Zhong, R.; Du, X.; and Du, Y. 2022. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–19.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Liu, M.; Chai, Z.; Deng, H.; and Liu, R. 2022. A CNN-Transformer Network With Multiscale Context Aggregation for Fine-Grained Cropland Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 15: 4297–4306.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*, 10012–10022.

Lv, L.; and Zhang, L. 2025. ScaleMatch: Multi-scale Consistency Enhancement for Semi-supervised Semantic Segmentation. In *AAAI*, volume 39, 5910–5918.

Mao, J.; Shen, Y.; Guo, J.; Yao, Y.; Hua, X.; and Shen, H. 2025. Prune and merge: Efficient token compression for vision transformer with spatial information preserved. *IEEE Trans. Multimed.*, 27: 4670–4683.

Noman, M.; Fiaz, M.; and Cholakkal, H. 2024. ChangeBind: A Hybrid Change Encoder for Remote Sensing Change Detection. In *IGARSS*, 8417–8422. IEEE.

Pei, G.; Chen, T.; Jiang, X.; Liu, H.; Sun, Z.; and Yao, Y. 2024. Videomac: Video masked autoencoders meet convnets. In *CVPR*, 22733–22743.

Pei, G.; Shen, F.; Yao, Y.; Xie, G.-S.; Tang, Z.; and Tang, J. 2022. Hierarchical feature alignment network for unsupervised video object segmentation. In *ECCV*, 596–613. Springer.

Pei, G.; and Zhang, L. 2022. Feature Hierarchical Differentiation for Remote Sensing Image Change Detection. *IEEE Geosci. Remote Sens. Lett.*, 19: 1–5.

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.; and Zhu, R.-J. 2023. RWKV: Reinventing RNNs for the Transformer Era. In *Findings of ACL: EMNLP 2023*, 14048–14077.

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.; and Zhu, R.-J. 2024. Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence. In *CoLM*.

Peng, B.; Zhang, R.; Goldstein, D.; Alcaide, E.; Du, X.; Hou, H.; Lin, J.; Liu, J.; Lu, J.; Merrill, W.; et al. 2025. RWKV-7 "Goose" with Expressive Dynamic State Evolution. *arXiv preprint arXiv:2503.14456*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 234–241. Springer.

Shen, L.; Lu, Y.; Chen, H.; Wei, H.; Xie, D.; Yue, J.; Chen, R.; Lv, S.; and Jiang, B. 2021. S2Looking: A Satellite Side-Looking Dataset for Building Change Detection. *Remote Sens.*, 13(24): 5094.

Sheng, M.; Sun, Z.; Cai, Z.; Chen, T.; Zhou, Y.; and Yao, Y. 2024a. Adaptive integration of partial label learning and negative learning for enhanced noisy label learning. In *AAAI*, volume 38, 4820–4828.

Sheng, M.; Sun, Z.; Pei, G.; Chen, T.; Luo, H.; and Yao, Y. 2024b. Enhancing Robustness in Learning with Noisy Labels: An Asymmetric Co-Training Approach. In *ACM MM*, 4406–4415.

Song, Z.; Wei, X.; Kang, X.; Li, S.; and Liu, J. 2023. Toward Efficient Remote Sensing Image Change Detection via Cross-Temporal Context Learning. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–10.

Su, Y.; Ma, P.; Wang, W.; Wang, S.; Wu, Y.; Li, Y.; and Jing, P. 2025. AMDANet: Augmented Multi-scale Difference Aggregation Network for Image Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–12.

Tang, X.; Zhang, T.; Ma, J.; Zhang, X.; Liu, F.; and Jiao, L. 2023. WNet: W-Shaped Hierarchical Network for Remote-Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–14.

Wang, J.; Song, J.; Zhang, H.; Zhang, Z.; Ji, Y.; Zhang, W.; Zhang, J.; and Wang, X. 2025. SPMNet: A Siamese Pyramid Mamba Network for Very-High-Resolution Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–14.

Wang, L.; Wang, L.; Wang, Q.; and Atkinson, P. M. 2022. SSA-SiamNet: Spectral-Spatial-Wise Attention-Based Siamese Network for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–18.

Wang, W.; Liu, C.; Liu, G.; and Wang, X. 2024. CF-GCN: Graph Convolutional Network for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–13.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *ECCV*, 3–19.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *NeurIPS*, 34: 12077–12090.

Xing, Y.; Jia, Y.; Gao, S.; Hu, J.; and Huang, R. 2025. Frequency-Enhanced Mamba for Remote Sensing Change Detection. *IEEE Geosci. Remote Sens. Lett.*, 22: 1–5.

Xu, J.; Luo, C.; Chen, X.; Wei, S.; and Luo, Y. 2021. Remote Sensing Change Detection Based on Multidirectional Adaptive Feature Fusion and Perceptual Similarity. *Remote Sens.*, 13: 3053.

Yang, F.; Li, M.; Shu, W.; Qin, A.; Song, T.; Gao, C.; and Xia, G.-S. 2025. ConvFormer-CD: Hybrid CNN-Transformer with Temporal Attention for Detecting Changes in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–15.

Ye, Y.; Wang, M.; Zhou, L.; Lei, G.; Fan, J.; and Qin, Y. 2023. Adjacent-Level Feature Cross-Fusion With 3-D CNN for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–14.

Zang, Q.; Yang, J.; Wang, S.; Zhao, D.; Yi, W.; and Zhong, Z. 2025. ChangeDiff: A Multi-Temporal Change Detection Data Generator with Flexible Text Prompts via Diffusion Model. In *AAAI*, volume 39, 9763–9771.

Zhang, C.; Wang, L.; Cheng, S.; and Li, Y. 2022. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 60: 1–13.

Zhang, H.; Chen, H.; Zhou, C.; Chen, K.; Liu, C.; Zou, Z.; and Shi, Z. 2024. BiFA: Remote Sensing Image Change Detection With Bitemporal Feature Alignment. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–17.

Zhang, H.; Chen, K.; Liu, C.; Chen, H.; Zou, Z.; and Shi, Z. 2025a. CDMamba: Incorporating Local Clues Into Mamba for Remote Sensing Image Binary Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–16.

Zhang, H.; Teng, Y.; Li, H.; and Wang, Z. 2025b. STRobustNet: Efficient Change Detection via Spatial-Temporal Robust Representations in Remote Sensing. *IEEE Trans. Geosci. Remote Sens.*, 63: 1–15.

Zhang, K.; Zhao, X.; Zhang, F.; Ding, L.; Sun, J.; and Bruzzone, L. 2023. Relation Changes Matter: Cross-Temporal Difference Transformer for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.*, 61: 1–15.

Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; and Ouyang, W. 2024. RS-Mamba for Large Remote Sensing Image Dense Prediction. *IEEE Trans. Geosci. Remote Sens.*, 62: 1–14.

Zheng, Z.; Ma, A.; Zhang, L.; and Zhong, Y. 2021. Change is Everywhere: Single-temporal Supervised Object Change Detection in Remote Sensing Imagery. In *ICCV*, 15193–15202.

Zhou, B.; Li, L.; Wang, Y.; Liu, H.; Yao, Y.; and Wang, W. 2025. UNIALIGN: Scaling Multimodal Alignment within One Unified Model. In *CVPR*, 29644–29655.

Zhou, X.; and Chen, T. 2024. BSBP-RWKV: Background Suppression with Boundary Preservation for Efficient Medical Image Segmentation. In *ACM MM*, 4938–4946.