Efficient-CANet: Rethinking Cross-Temporal Interaction and Multi-Scale Geometry for Ultra-Lightweight Change Detection

Xingpeng Zhang [1], Dian Qi [1], Yuru Li [1], Xingpan Hu [1], Qiuli Wang [23], Yang Yu †[1]
[1]School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu, 610500 China
[2]Department of Radiology, The First Affiliated Hospital of Army Medical University, Chongqing 400032, China
[3]Yu-Yue Pathology Research Center, Jinfeng Laboratory, Chongqing 400032, China

***Abstract*—Building change detection in complex urban environments faces a dilemma: distinguishing genuine changes from pseudo-changes (e.g., illumination, seasonal variations) requires robust global context modeling, yet deployment on resource-constrained edge devices demands extreme efficiency. Existing lightweight methods often compromise discriminative capability or multi-scale adaptability to reduce parameters. To address this trade-off, we propose CANet, an ultra-lightweight network (0.73M parameters) designed for efficient and accurate change detection. CANet introduces two core innovations: (1) A Cross-Temporal Change-Aware Module (CTCAM), which incorporates a Factorized Similarity Kernel (FSK). By factorizing the global correlation matrix, FSK explicitly models inter-phase dependencies with linear complexity $O(N)$, effectively suppressing pseudo-change noise without the computational burden of standard self-attention. (2) An Omni-Dimensional Spatial Pyramid (ODSP), which synergizes dynamic kernel weighting with multi-scale dilated convolutions. This design decouples geometric deformation from scale variations, enabling adaptive feature extraction for buildings of diverse shapes and sizes. Extensive experiments on the LEVIR-CD and S2Looking datasets demonstrate that CANet achieves SOTA accuracy among lightweight models, surpassing the heavier ChangerEx while using 16× fewer parameters.**

***Index Terms*—Change detection, lightweight networks, cross-temporal interaction, dynamic multi-scale fusion.**

I. Introduction

Urbanization accelerates the dynamic evolution of cityscapes, making automated building change detection (CD) a critical task for urban planning and land management. While early methods relied on handcrafted features [1], the field has been revolutionized by Deep Learning. Contemporary approaches, ranging from Convolutional Neural Networks (CNNs) to Vision Transformers (ViTs) and recent Mamba-based architectures, have significantly improved detection accuracy by leveraging deep contextual representations [2]. CNN-based methods utilize twin architectures and attention mechanisms to generate multi-scale feature differences and extract local contexts; however, they remain limited in modeling global features [3]–[7]. In contrast, ViT-based methods capture long-range dependencies through global self-attention [8], [9], enabling comprehensive spatio-temporal modeling, as demonstrated by STCD [10], LSAT [11], and MISGNet [12]. More recently, Mamba-based architectures [13], [14] have been introduced to further advance global feature extraction in change detection tasks.

However, the deployment of these advanced models on edge devices (e.g., UAVs) is hindered by their high computational demands. Although Transformer-based and Mamba-based methods excel at capturing long-range dependencies necessary for distinguishing real changes from environmental noise, their quadratic or high linear computational complexity renders them impractical for real-time applications. Conversely, existing lightweight networks often achieve efficiency by simplifying the interaction mechanisms of features via three strategies: (1) architecture optimization, as seen in TinyCD's compact U-Net with hybrid attention [15], LightCDNet's early fusion and pyramid decoding [16], LWCCNet's feature pyramid reconstruction [17], and LFHNet's use of dilated convolution and cross-head attention [18]; (2) Transformer compression, exemplified by LWT's use of depthwise separable convolution to reduce key-value dimensions [19] and other approaches that simplify attention mechanisms [20]; and (3) knowledge distillation, which transfers capabilities from large teachers to compact student models for better efficiency-accuracy trade-offs [21], [22]. This reductionism typically leads to two critical limitations: (1) Insufficient Cross-Temporal Alignment. Simplified attention mechanisms struggle to explicitly model the correlation between bi-temporal images, failing to filter out illumination differences or seasonal shifts effectively. (2) Rigid Geometric Modeling. Standard convolutions with fixed kernels lack the flexibility to adapt to buildings with extreme scale variations and irregular geometries, which are common in high-resolution remote sensing imagery.

To bridge the gap between high performance and deployment efficiency, this paper proposes CANet, a lightweight change detection framework that rethinks cross-temporal interaction and multi-scale adaptation. Our design philosophy focuses on removing redundant computations while maximizing the representation of "change-relevant" features.

This work is supported by the National Natural Science Foundation of China (No. 62441610 and U23A2046) and the Chongqing Science and Health Joint Medical Research Project (No. 2025MSXM016). † Corresponding author: yuyang@swpu.edu.cn

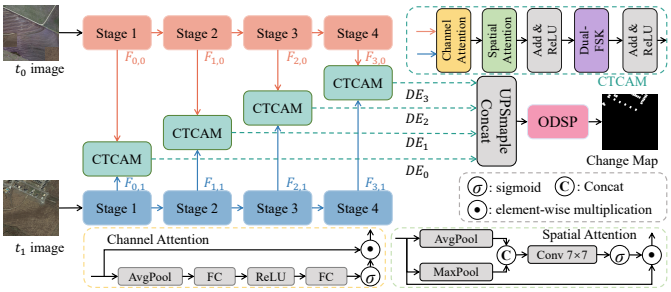

Fig. 1. The overall framework of CANet.

Specifically, we introduce two key innovations. First, to address the alignment challenge without incurring quadratic costs, we propose the Cross-Temporal Change-Aware Module (CTCAM) based on a Factorized Similarity Kernel (FSK). Unlike standard self-attention that computes a dense $N \times N$ correlation map, FSK factorizes the attention operation to achieve linear complexity $O(N)$. This allows the model to capture global inter-phase dependencies and suppress pseudo-changes efficiently. Second, to tackle geometric and scale variations, we design the Omni-Dimensional Spatial Pyramid (ODSP). While traditional ASPP handles scale via dilated convolutions and ODConv handles geometry via dynamic weights, they address these problems in isolation. ODSP unifies these concepts, employing dynamic kernel weighting across spatial, channel, and kernel dimensions within a multi-scale pyramid structure. This synergy enables the network to adaptively "deform" its effective receptive field, ensuring precise boundary delineation for buildings of varying sizes.

- We propose a Cross-Temporal Change-Aware Module (CTCAM) that incorporates Factorized Similarity Kernels to model inter-phase correlations with linear complexity, thereby reducing computational cost while enhancing discriminative representation.
- We design an Omni-Dimensional Spatial Pyramid (ODSP) that integrates dynamic kernel weighting with multi-scale dilated convolutions, effectively decoupling geometric distortions from scale variations for robust building detection.
- We present CANet, an ultra-lightweight architecture (0.73M parameters) based on a truncated EfficientNet-B4 backbone. Extensive experiments on LEVIR-CD and S2Looking datasets show that CANet achieves SOTA performance among lightweight and heavy models.

II. Methodology

This section details the architecture of CANet, a lightweight change detection model featuring a twin backbone, CTCAM for cross-temporal modeling, and ODSP for multi-scale fusion. The overall framework is illustrated in Fig. 1.

A. Efficiency-Oriented Design Choices

Despite using EfficientNet-B4 as the backbone, we adopt only its first four blocks and replace standard convolutions with depthwise separable convolutions wherever possible. This truncated and factorized design significantly reduces the number of parameters while preserving multi-scale feature extraction capabilities. Furthermore, we introduce Factorized Similarity Kernels (FSK) to replace quadratic self-attention, reducing the complexity from $O(N^2)$ to $O(N)$. These choices collectively enable CANet to maintain a lightweight profile without sacrificing representational power.

Following Tiny-CD's configuration and ablation results [15], we adopt the first four blocks of EfficientNet-B4 [23] as the shared-parameter backbone to extract multi-stage features from dual-phase inputs t_0 and t_1, i.e. $F_{i,j} = backbone(t_j)$, where $F_{i,j} \in \mathbb{R}^{H_i \times W_i \times C_i}$ are the learned features of the i stage under the j-th phase, $i = 0, 1, 2, 3$ represents the encoder stage, and $j = 0, 1$ represents two time dimensions.

B. Cross-Temporal Change-Aware Module

Inspired by cross-modal learning, where bi-temporal remote sensing images with significant variations can be treated as distinct "modalities", we design a Cross-Temporal Change-Aware Module (CTCAM) to efficiently capture inter-phase correlations, as illustrated in Fig. 2. To mitigate the high computational cost of standard attention mechanisms, CTCAM incorporates a Factorized Similarity Kernel (FSK), which reduces complexity while enhancing discriminative representation. This module uses the differential information of dual-time-phase characteristics as its core input, mapping each branch to the query and value components, respectively. By leveraging the properties of the attention mechanism, it ultimately achieves targeted and efficient enhancement of differentiated information.

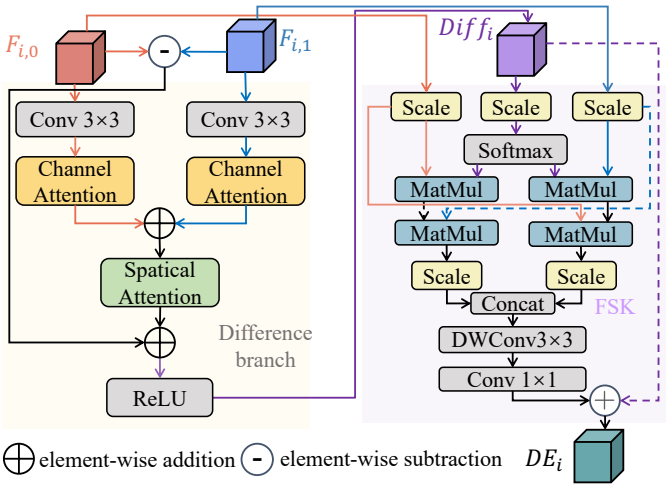

Fig. 2. The structure of the CTCAM module.

Step 1. Enhance difference via channel-spatial attention.

First, we compute the absolute difference between the bi-temporal features $D_i = \|F_{i,0} - F_{i,1}\|$. Then, we apply a

lightweight channel-spatial attention mechanism to enhance discriminative features.

$$F'_{i,j} = Conv_{3\times3}(F_{i,j}) \tag{1}$$
$$CA_{i,j} = F'_{i,j} \odot \sigma(FC_2(FC_1(AvgP(F'_{i,j}))))$$

where FC_1 reduces channels to $\frac{C_i}{8}$ and FC_3 restores the original dimension, $\odot$ denotes element-wise multiplication, and σ is the sigmoid function. The combined channel attention is obtained by $CA_i = CA_{i,0} + CA_{i,1}$.

To capture a larger receptive field, a 7×7-sized convolution kernel is adopted here. Spatial attention is then applied:

$$SA_i = CA_i \odot \sigma(Conv_{7\times7}(Concat(AvgP(CA_i), MaxP(CA_i))))$$

The enhanced difference feature is computed as:

$$Diff_i = ReLU(SA_i \oplus D_i) \tag{2}$$

Step 2. Cross-Temporal Feature Modulation with FSK

For feature group $c^l_{(i,j)}$ divided into G groups along the channel dimension, we compute:

$$s^l_{(i,j)} = (Softmax(k^l_{(i,j)}))^T \cdot v^l_{(i,j)}$$
$$D^{fs}_l(c^l_{(i,j)}) = q^l_{(i,j)} \cdot s^l_{(i,j)} \tag{3}$$

The modulated features from both temporal branches are concatenated and fused.

$$S_i = Concat(FSK(F_{i,0}, Diff_i, F_{i,1}), FSK(F_{i,1}, Diff_i, F_{i,0}))$$
$$DS_i = ReLU(BN(DWConv_{3\times3}(S_i))) \tag{4}$$

Mechanism Analysis. Drawing inspiration from Linear Transformers [24], FSK adapts the linearized attention mechanism to the bi-temporal domain. We argue that the quadratic complexity ($O(N^2)$) of standard self-attention is redundant for change detection. This is because valid changes are typically sparse, and distinguishing them relies primarily on contrasting local features against a global semantic context rather than establishing dense pixel-to-pixel dependencies. FSK addresses this by formulating a low-rank approximation of the global attention map.

Specifically, the input feature $C_{(i,j)}$ is first divided into L groups. For each group, the key component $k^l_{(i,j)}$ undergoes softmax normalization to generate a global similarity distribution. This distribution is then used to aggregate the value component $v^l_{(i,j)}$, resulting in the Factorized Similarity Kernel (FSK) $s^l_{(i,j)}$. This kernel encapsulates global contextual dependencies in a compact form. By interacting the query component $q^l_{(i,j)}$ with this pre-computed kernel, global semantic information is efficiently distributed to local features to derive the enhanced representation $D^{fs}_l(c^l_{(i,j)})$. Consequently, as illustrated in Fig. 3, by exploiting the associative property of matrix multiplication, the computational complexity is reduced from $O(N^2)$ to linear $O(N)$.

Finally, we obtain the enhanced differential feature $DE_i = Conv(DS_i, 1) + Diff_i$. This design effectively balances the need for cross-temporal modeling with the efficiency requirements of edge deployment.

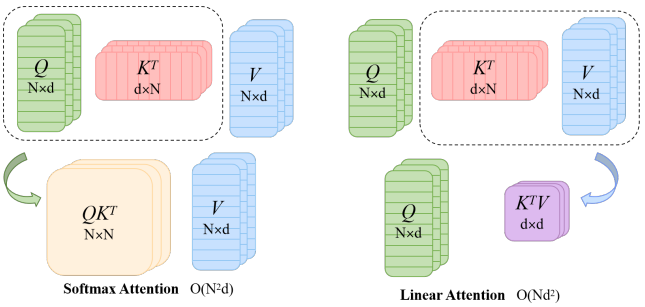

Fig. 3. FSK achieves linear self-attention

C. Omni-Dimensional Spatial Pyramid

Building change detection requires managing structures of varying scales and shapes. While ASPP [25] captures multi-scale context using parallel dilated convolutions, it relies on fixed kernels that lack adaptability to diverse input characteristics. In contrast, ODConv [26] introduces dynamic weighting across four dimensions but functions at a single scale. Our ODSP module, illustrated in Fig. 4, synergistically combines the strengths of both approaches. First, from ASPP, we adopt the multi-scale paradigm using parallel branches with dilation rates $\{1, 6, 12, 18\}$. Second, from ODConv, we incorporate dynamic kernel weighting across spatial, input-channel, output-channel, and kernel dimensions. Third, our innovation lies in integrating these concepts into a unified pyramid structure that adaptively captures multi-scale contextual information.

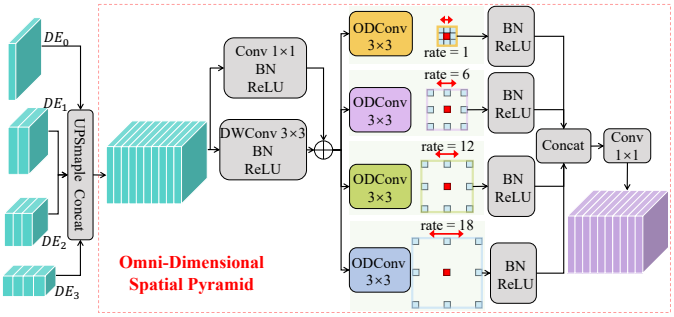

Fig. 4. Overview of ODSP module after upsampling and concat operation.

After upsampling and concatenating the enhanced differential features DE_i obtained in the previous section, depth-separable convolution was initially employed to get the fused feature $FP_{in} \in \mathbb{R}^{H\times W\times C}$. Then, the ODSP output is computed as $FP^{out}_i = ReLU(BN(ODConv(FP_{in}, k_i, r_i)))$, where $k_i \in \{1, 3, 3, 4\}$ and $r_i \in \{1, 6, 12, 18\}$ denote kernel sizes and dilation rates respectively. The final output integrates multi-scale representations:

$$FP^{out} = Conv(FP^{out}_0, FP^{out}_1, FP^{out}_2, FP^{out}_3, 1) \tag{5}$$

Unlike standard ASPP, ODSP employs dynamic kernels. Compared to standalone ODConv, our pyramid structure captures multi-scale context essential for building change detection. This hybrid design enables the model to simultaneously

TABLE I
EXPERIMENTAL RESULTS OF THE LEVIR-CD DATASET. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE, AND GREEN RESPECTIVELY.

methods	BackBone	Params(M)	Precision(%)	Recall(%)	F1(%)	IoU(%)	OA(%)
DTCDSCN [27]	SE-Res34	41.07	88.53	86.83	87.67	78.05	98.77
STANet [4]	ResNet18	12.21	83.81	90.00	87.26	77.40	98.66
CDNet [28]	ResNet18	14.33	91.60	86.50	89.00	-	-
IFN* [29]	VGG-16	36.00	91.17	90.51	90.83	84.98	99.09
ChangeFormer* [9]	MiT-b1	13.94	92.59	89.68	91.11	83.66	99.12
ChangerAlign [3]	ResNet18	11.39	93.30	89.59	91.41	84.24	99.14
ChangerEx [3]	ResNet18	11.39	92.97	90.61	91.77	84.86	99.17
ELGC-Net [5]	VGG-16	10.57	-	-	91.2	83.83	99.12
MAMFANet [6]	ResNet34	76.81	93.05	89.91	91.45	84.24	-
CDMamba [14]	ConvMamba	11.9	91.43	90.08	90.75	83.07	99.06
MISGNet [12]	PVTv2-B1	15.45	92.75	90.69	91.7	84.69	-
LTMFFNet [19]	Swin Transformer	17.64	92.4	90.23	91.30	83.89	99.53
FC-EF [30]	-	1.35	86.91	80.17	83.40	71.53	98.39
FC-Siam-Diff [30]	-	1.35	89.53	83.31	86.31	75.92	98.67
BiT* [8]	ResNet18	2.99	91.97	88.62	90.26	80.68	98.92
SNUNet [31]	-	3.01	92.45	90.17	91.30	84.02	99.27
LightCDNet-small [16]	ShuffleNetV2	0.35	91.36	89.81	90.57	82.77	99.07
LightCDNet-base [16]	ShuffleNetV2	1.32	92.12	90.43	91.27	83.94	99.12
LightCDNet-large [16]	ShuffleNetV2	2.82	92.43	90.45	91.43	84.21	99.14
BiFA [22]	SegformerB0	3.8	91.52	89.86	90.69	82.96	99.06
FTAN [7]	-	-	92.41	88.82	90.51	82.78	-
TinyCD* [15]	EfficientNet-B4	0.28	92.68	89.47	91.05	83.57	99.09
LFHNet [18]	-	3.78	92.75	87.92	90.27	82.27	99.03
CANet*	EfficientNet-B4	0.73	92.81	90.81	91.80	84.84	99.17

address scale variation and feature adaptability—two critical challenges in change detection.

III. EXPERIMENTS

A. Datasets and Setting

This paper utilizes the commonly used datasets LEVIR-CD [4] and S2Looking [32] to evaluate the model's performance. We employ five evaluation metrics to assess change detection results: Precision (Pre), Recall (Re), F1-score (F1), Intersection over Union (IoU), and Overall Accuracy (OA). All experiments were conducted using PyTorch and OpenMMLab on an NVIDIA RTX 3090 GPU. The model was pre-trained on ImageNet-1k and fine-tuned using the AdamW optimizer (weight decay of 0.05) with a polynomially decaying learning rate starting at 0.005. Training was performed with a batch size of 16 and an input size of 256×256 pixels, with 160k and 80k iterations for S2Looking and LEVIR-CD, respectively.

On the training set, we conducted preprocessing and data augmentation operations. Specifically, we implemented the following techniques: normalization; flipping the image with a 50% probability; rotating the image with a 50% probability within an angle range of -20° to 20°; swapping the sequence of dual-phase images with a 50% probability; and applying photometric transformations to adjust the image's brightness, contrast, saturation, and hue, also with a 50% probability.

B. Experimental results of the LEVIR-CD dataset

Quantitative Analysis.[1] As shown in Table I, CANet achieves Precision, Recall, F1, IoU, and OA of 92.81%, 90.81%, 91.80%, 84.84%, and 99.17%, respectively. Compared to TinyCD with the same backbone, CANet improves F1 by 0.75% and IoU by 1.27%. It also surpasses LightCDNet-large by 0.37% in F1 and 0.63% in IoU despite having comparable or fewer parameters. Notably, CANet outperforms the much larger ChangeEx model in Recall and F1 scores while using only 1/16 of its parameters, demonstrating a favorable accuracy-efficiency trade-off.

Qualitative Analysis. As shown in Fig. 5, qualitative comparisons on LEVIR-CD demonstrate CANet's effectiveness across varying building scales. In scenes containing both large and small structures (Fig. 5(b)), CANet accurately delineates building contours and captures fine-scale details with minimal omission or false detection. Moreover, under challenging conditions where building colors resemble surrounding roads (Fig. 5(c)), CANet maintains reliable detection performance, outperforming other lightweight models that struggle with such color ambiguities. These results highlight CANet's robustness in handling scale variation and color similarity, attributed to its adaptive feature extraction and cross-temporal modeling mechanisms.

C. Experimental results of the S2Looking dataset

Quantitative Analysis. As shown in Table II, the proposed CANet achieves Precision, Recall, F1, IoU, and OA of 72.78%, 59.59%, 65.53%, 48.73%, and 99.26%, respectively, on the S2Looking dataset. Compared to TinyCD with the same backbone, CANet improves F1 by 2.17% and IoU by 2.36% with only a marginal parameter increase of 0.45M. Against LightCDNet-large, CANet attains higher F1 (+3.6%) and IoU (+3.95%) despite using fewer parameters. Overall, CANet outperforms existing lightweight models across most metrics, with a parameter count only slightly above TinyCD and LightCDNet-small. Notably, it matches or exceeds the per-

[1]In Table I and II, * indicates the use of a pre-trained model.

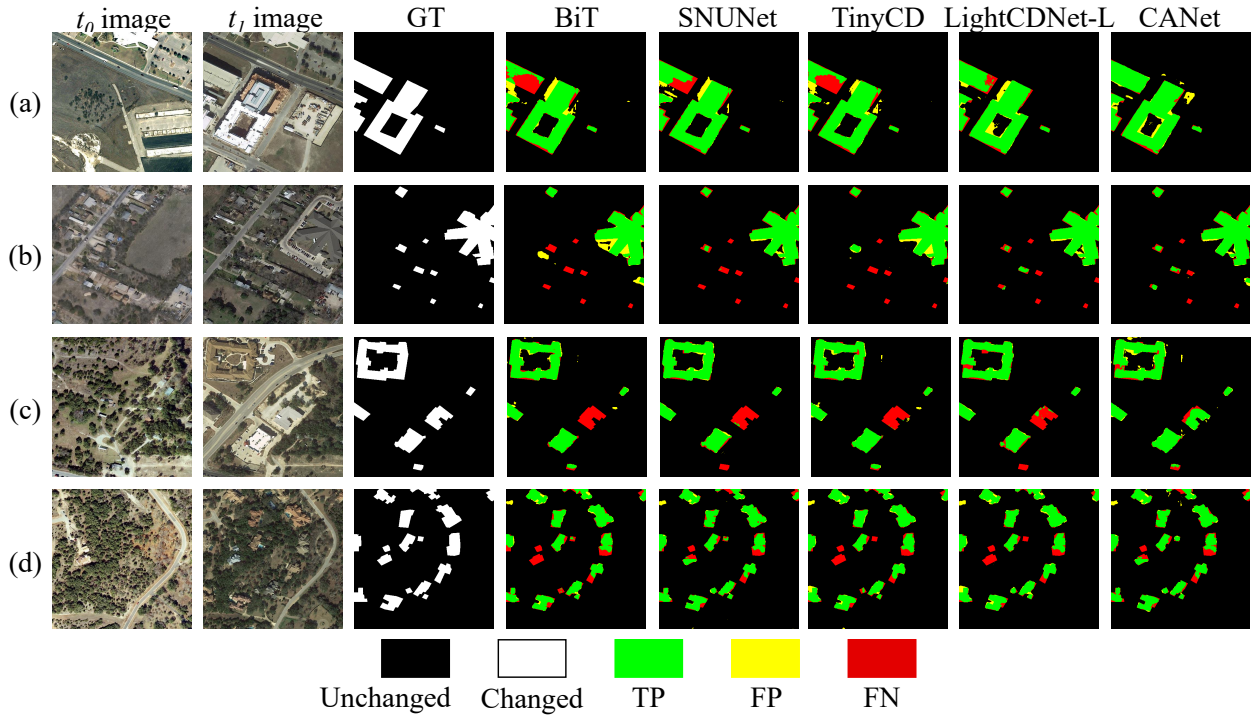

Fig. 5. Visualization results of the LEVIR-CD dataset

TABLE II
EXPERIMENTAL RESULTS OF THE S2LOOKING DATASET. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE, AND GREEN RESPECTIVELY.

methods	BackBone	Params(M)	Precision(%)	Recall(%)	F1(%)	IoU(%)	OA(%)
DTCDSCN [27]	SE-Res34	41.07	68.58	49.16	57.27	40.12	-
STANet [4]	ResNet18	12.21	38.75	56.49	45.97	29.84	98.82
CDNet [28]	ResNet18	14.33	67.48	54.93	60.56	43.43	-
IFN* [29]	VGG-16	36.00	66.46	61.95	64.13	47.39	99.16
ChangeFormer* [9]	MiT-b1	13.94	72.82	56.13	63.39	46.73	99.06
ChangerVanilla [3]	ResNet18	11.39	72.59	58.25	64.63	47.73	99.23
ChangerAlign [3]	ResNet18	11.39	71.62	60.06	65.33	48.31	99.24
ChangerEx [3]	ResNet18	11.39	73.59	60.15	66.20	49.57	99.27
FC-EF [30]	-	1.35	81.36	8.95	7.65	8.77	-
FC-Siam-Conc [30]	-	1.54	68.27	18.52	13.54	17.05	-
FC-Siam-Diff [30]	-	1.35	83.29	15.76	13.19	15.28	-
BiT* [8]	ResNet18	2.99	74.80	55.56	63.76	44.77	98.57
SNUNet [31]	-	3.01	71.94	56.34	63.19	46.24	99.11
LightCDNet-small [16]	ShuffleNetV2	0.35	64.73	52.93	58.24	41.08	98.95
LightCDNet-base [16]	ShuffleNetV2	1.32	63.04	57.36	60.07	42.93	99.08
LightCDNet-large [16]	ShuffleNetV2	2.82	62.97	60.79	61.86	44.78	99.09
TinyCD* [15]	EfficientNet-B4	0.28	68.12	59.22	63.36	46.37	99.14
CANet*	EfficientNet-B4	0.73	72.78	59.59	65.53	48.73	99.27

formance of significantly larger models such as ChangerAlign and ChangerEx while using merely 1/16 of their parameters.

Qualitative Analysis. Fig. 6 presents a visual comparison between CANet and other lightweight models on the S2Looking dataset. In densely packed small-scale building areas (Fig. 6(a)), CANet achieves a favorable balance between false positives and false negatives despite the challenging proximity and size of structures. For large-scale buildings (Fig. 6(c)-(d)), CANet demonstrates a lower false detection rate and produces boundaries that align more accurately with ground truth, confirming its robustness across varying building scales.

D. *Efficiency Analysis*

Although parameter count is a commonly used lightweight metric, we also provide estimated FLOPs and inference speed for a more comprehensive efficiency assessment. CANet requires approximately 3.2 GFLOPs for a 256×256 input, which is comparable to TinyCD (3.0 GFLOPs) and significantly lower than ChangerEx (approximately 50 GFLOPs). Additionally, CANet achieves around 22 FPS, demonstrating its suitability for real-time edge deployment.

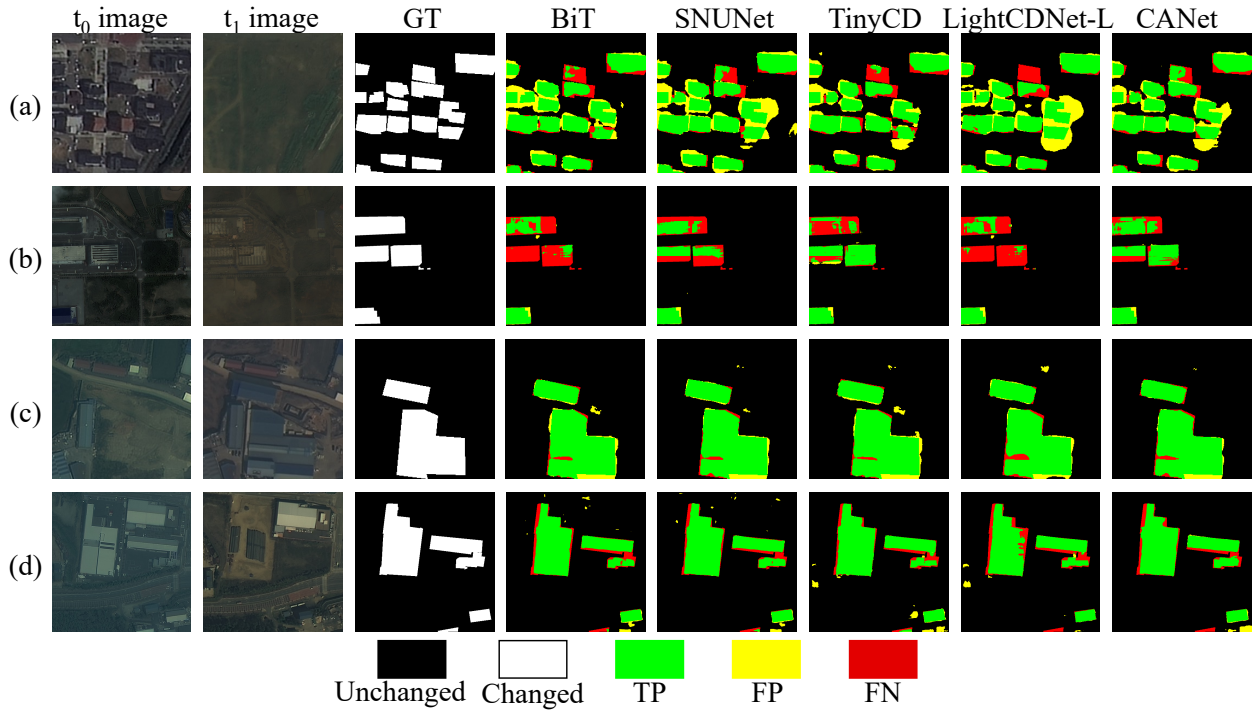

Fig. 6. Visualization results of the S2Looking dataset

E. Ablation experiment

Ablation studies on the S2Looking dataset evaluate module effectiveness using F1 and IoU metrics, with all models trained consistently under BCE Loss.

TABLE III
MODULE ANALYSIS VALIDITY VERIFICATION EXPERIMENT

CTCAM	ODSP	Pre(%)	Re(%)	F1(%)	IoU(%)	OA(%)
-	-	63.04	57.36	60.07	42.93	99.08
✓	-	70.81	58.22	63.9	46.95	99.20
-	✓	68.54	58.61	62.60	45.36	99.15
✓	✓	**72.78**	**59.59**	**65.53**	**48.73**	**99.27**

Module Analysis. Ablation studies in Table III confirm the complementary synergy of CTCAM and ODSP. While CTCAM alone improves the baseline (60.07% F1, 42.93% IoU) to 63.9% F1 and 46.95% IoU through differential enhancement and cross-temporal modeling, and ODSP alone achieves 63.60% F1 and 46.36% IoU via multi-scale dynamic convolution, their integration in CANet yields optimal performance (65.53% F1, 48.73% IoU), surpassing both the baseline and the sum of individual gains.

Ablation for CTCAM. As shown in Table IV, our channel-spatial difference enhancement (Diff Only) improves F1 by 2.08% and IoU by 2.15% over direct subtraction, while bidirectional attention brings additional gains of 1.12% F1 and 1.18% IoU over unidirectional attention. The complete CTCAM integrating both mechanisms achieves optimal performance (63.9% F1, 46.95% IoU), demonstrating the complementary benefits of enhanced difference representation and bidirectional temporal modeling.

TABLE IV
CTCAM VALIDITY VERIFICATION EXPERIMENT

methods	Pre(%)	Re(%)	F1(%)	IoU(%)	OA(%)
Sub	63.04	57.36	60.07	42.93	99.08
Sum	70.19	49.33	58.27	41.12	99.14
Concat	69.30	52.26	59.59	42.44	99.14
Diff Only	71.0	55.26	62.15	45.08	99.19
Single FSK	64.61	59.36	61.87	44.79	99.11
Double FSK	70.37	57.01	62.99	45.97	99.19
Double FSK with Diff	66.23	**60.16**	63.60	46.63	99.15
CTCAM	**72.78**	59.59	**65.53**	**48.73**	**99.27**

TABLE V
ODSP VALIDITY VERIFICATION EXPERIMENT

methods	Params(M)	Pre(%)	Re(%)	F1(%)	IoU(%)
ASPP (+GAP)	1.23	71.57	59.02	64.69	47.81
with DWConv	**0.71**	71.49	58.84	64.55	47.66
with ODConv	0.75	72.39	59.27	65.18	48.34
ODSP (w.o. GAP)	0.73	**72.78**	**59.59**	**65.53**	**48.73**

Ablation for ODSP: Synergy of Scale and Geometry. Table V confirms that ODSP's efficacy stems from structural synergy rather than parameter growth. Replacing static DWConv with ODConv yields gains of 0.63% F1 and 0.68% IoU. This indicates that while dilated convolutions manage *scale variations*, dynamic ODConv is crucial for handling *geometric deformations*, effectively decoupling these two factors. Furthermore, removing the Global Average Pooling (GAP) branch boosts F1

to 65.53%. Unlike semantic segmentation, fine-grained change detection benefits more from preserving local high-frequency details than aggregating global context, which often introduces environmental noise.

F. Limitations and future works

Despite its effectiveness, CANet has limitations in robustness under complex interference (e.g., rain, haze, occlusion) and generalization across modalities (e.g., SAR, LiDAR) or multi-temporal sequences. Future work will focus on enhancing robustness through ODConv optimization and lightweight super-resolution branches, improving generalization via modality-adaptive backbones and temporal modeling, and developing cross-modal fusion strategies to distinguish real changes from pseudo-changes.

IV. Conclusion

This paper proposes CANet, a lightweight change detection model based on EfficientNet-B4, which introduces a Cross-phase Change Perception Module (CTCAM) to enhance differential features and model temporal correlations, as well as an Omni-Dimensional Spatial Pyramid (ODSP) module that dynamically adapts to multi-scale buildings. Evaluated on the LEVIR-CD and S2Looking datasets, CANet achieves competitive accuracy with only 0.73M parameters, effectively balancing efficiency and performance. Additionally, ablation studies validate the contribution of each module.

References

[1] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based markov random field model," *IEEE TIP*, vol. 29, pp. 757–767, 2020.

[2] X. Zhang, Y. Li, Q. Wang, and S. Wu, "Mfi-cd: a lightweight siamese network with multidimensional feature interaction for change detection," *Int. J. Remote Sens.*, vol. 45, no. 8, pp. 2548–2566, 2024.

[3] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE TGRS*, vol. 61, pp. 1–11, 2023.

[4] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.

[5] M. Noman, M. Fiaz, H. Cholakkal, S. H. Khan, and F. S. Khan, "Elgc-net: Efficient local-global context aggregation for remote sensing change detection," *IEEE TGRS*, vol. 62, pp. 1–11, 2024.

[6] H. Ren, M. Xia, L. Weng, K. Hu, and H. Lin, "Dual-attention-guided multiscale feature aggregation network for remote sensing image change detection," *JSTARS*, vol. 17, pp. 4899–4916, 2024.

[7] C. Yu, H. Li, Y. Hu, Q. Zhang, M. Song, and Y. Wang, "Frequency-temporal attention network for remote sensing imagery change detection," *IEEE Geosci. Remote. Sens. Lett.*, vol. 21, pp. 1–5, 2024.

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

[9] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IEEE Int. Geosci. Remote Sens. Symposium*, 2022, pp. 207–210.

[10] D. Wang, X. Chen, N. Guo, H. Yi, and Y. Li, "Stcd: Efficient siamese transformers-based change detection method for remote sensing images," *Geo-Spat. Inf. Sci.*, vol. 27, no. 4, pp. 1192–1211, 2024.

[11] T. Lei, Y. Xu, H. Ning, Z. Lv, C. Min, Y. Jin, and A. K. Nandi, "Lightweight structure-aware transformer network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2023.

[12] B. Cui, C. Liu, H. Li, and J. Yu, "Misgnet: A multilevel intertemporal semantic guidance network for remote sensing images change detection," *JSTARS*, vol. 18, pp. 1827–1840, 2025.

[13] X. Zhang, S. Wu, Q. Wang, and K. Wang, "Effretmamba: A light-weight mamba for low-light image enhancement," *Circuits Syst. Signal Process.*, vol. 44, p. 8691–8711, 2025.

[14] H. Zhang, K. Chen, C. Liu, H. Chen, Z. Zou, and Z. Shi, "Cdmamba: Incorporating local clues into mamba for remote sensing image binary change detection," *IEEE TGRS*, vol. 63, pp. 1–16, 2025.

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

[16] Y. Xing, J. Jiang, J. Xiang, E. Yan, Y. Song, and D. Mo, "Lightcdnet: Lightweight change detection network based on VHR images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[17] L. Wang, D. Gao, X. Li, X. Tang, X. Zhao, and H. Gao, "Lwccnet: Lightweight remote sensing change detection network based on composite convolution operator," *JSTARS*, vol. 16, pp. 8621–8631, 2023.

[18] X. Jiang, S. Zhang, J. Gan, J. Wei, and Q. Luo, "Lfhnet: Lightweight full-scale hybrid network for remote sensing change detection," *JSTARS*, vol. 17, pp. 10 266–10 278, 2024.

[19] J. Li, P. Zheng, and L. Wang, "Remote sensing image change detection based on lightweight transformer and multi-scale feature fusion," *JSTARS*, vol. 18, pp. 5460–5473, 2025.

[20] L. Xue, X. Wang, Z. Wang, L. Zhang, and G. Li, "A lightweight patch-level change detection network via exploring the potential of pruning and multi-scale pooling," in *IEEE Int. Geosci. Remote Sens. Symposium*, 2024, pp. 8392–8396.

[21] G. Wang, N. Zhang, J. Wang, W. Liu, Y. Xie, and H. Chen, "Knowledge distillation-based lightweight change detection in high-resolution remote sensing imagery for on-board processing," *JSTARS*, vol. 17, pp. 3860–3877, 2024.

[22] H. Zhang, H. Chen, C. Zhou, K. Chen, C. Liu, Z. Zou, and Z. Shi, "Bifa: Remote sensing image change detection with bitemporal feature alignment," *IEEE TGRS*, vol. 62, pp. 1–17, 2024.

[23] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.

[24] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *ICML*, 2020, pp. 5156–5165.

[25] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 833–851.

[26] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," in *ICLR*, 2022.

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

[28] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE TGRS*, vol. 60, pp. 1–16, 2022.

[29] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS-J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, 2020.

[30] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *ICIP*, 2018, pp. 4063–4067.

[31] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[32] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang, "S2looking: A satellite side-looking dataset for building change detection," *Remote. Sens.*, vol. 13, no. 24, p. 5094, 2021.