Change3D: Revisiting Change Detection and Captioning from A Video Modeling Perspective

Duowang Zhu[1], Xiaohu Huang[2], Haiyan Huang[1], Hao Zhou[3], Zhenfeng Shao[1] ✉

[1]Wuhan University [2]The University of Hong Kong [3]Bytedance

Project Page

CVPR Nashville JUNE 11-15, 2025

Introduction

Background

As shown in Fig. 1 (a), existing change detection and captioning methods typically follow such a paradigm: (1) A pair of bi-temporal images are treated as distinct inputs, processing each image individually through a shared-weight image encoder for spatial feature extraction. (2) Then, a well-designed change extractor, primarily leveraging attention mechanisms, is introduced for bi-temporal feature interaction to detect the differences between them. (3) Finally, a decoder is employed to restore the feature map to its original resolution.

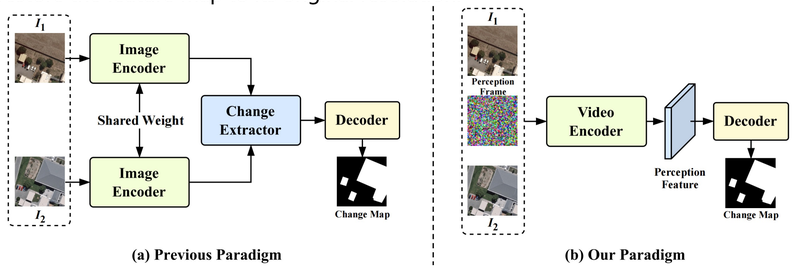

Figure 1. Previous paradigm *vs.* our paradigm.

Challenges

- **Inefficient Feature Encoding:** Task-agnostic image encoding fails to effectively focus on changes, leading to imbalanced parameter distribution with the majority dedicated to image encoding rather than change extraction (shown in Fig. 2).
- **Absence of Unified Framework:** Diverse change extractors tailored for various detection and captioning tasks complicate the creation of a unified framework.

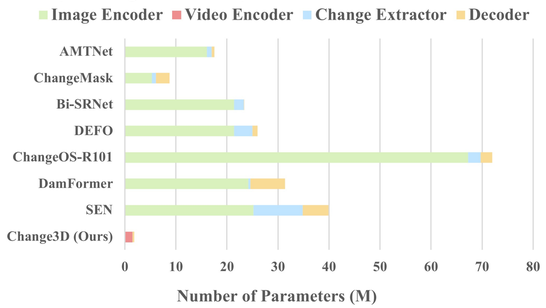

Figure 2. Parameter distribution of different methods.

Key Contributions

Motivated by **video modeling techniques that effectively capture relationships between images**, we redefine bi-temporal change detection and captioning tasks from a video modeling perspective (see in Fig. 1 (b)). We introduce **Change3D** with the following key contributions:

- **3D Video Modeling Paradigm:** Enable unified bi-temporal feature learning with joint spatial-temporal modeling, surpassing traditional 2D methods.
- **Efficient Dynamic Perception:** Utilize learnable perception frames for effective change extraction through direct feature interaction.
- **State-of-the-Art Performance:** Achieve superior results on **eight** benchmarks across **four** tasks while using only **6%–13%** of the parameters compared to current leading algorithms.

Method

To evaluate the multi-task adaptability of the Change3D framework, we apply it to remote sensing change detection and captioning tasks. As shown in Fig. 3, the core workflow includes:

Perception Feature Extraction

- **Perception Frame Initialization**: Dynamically generate learnable perception frames I_P based on the number of task head N.

  I_P = nn.Parameter(torch.randn(N, C, H, W)), using PyTorch code for expression

- **Spatiotemporal Input Construction**: Stack bi-temporal images and perception frames along the temporal dimension to form a 3D video-like sequence.

- **Perception Feature Learning**: A video encoder facilitates the interaction between the perception frames and bi-temporal images to produce perception features.

$$f = \mathcal{F}_{\text{enc}}(I_1 \copyright I_P \copyright I_2), \quad f \in \{\mathcal{R}^{T\times C_i\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}}\}_{i=0}^{3} \quad (T = N+2)$$

Perception feature for change detection: $p_{det} = f[1:1+N,:,:,:]$

Perception feature for change captioning: $p_{cap} = f[1,:,:,:]$

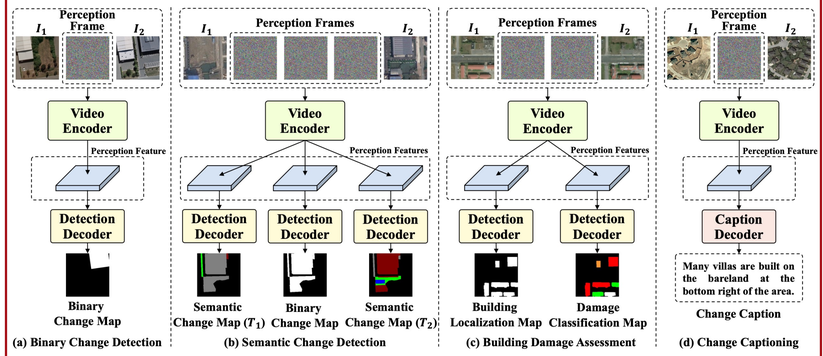

Figure 3. Unified framework for multiple change detection and captioning tasks.

Deocder and Optimization

- **Change Decoder:** To more convincingly highlight the learning prowess of Change3D, we choose to implement a simpler decoder, consisting of cascade convolutional layer followed by an upsampling operation.

$$p_{\text{det}}^{i,j} \leftarrow \text{Deconv}_{4\times4}(\text{Conv}_{1\times1}(p_{\text{det}}^{i+1,j})) + p_{\text{det}}^{i,j}$$

$$\mathcal{M} = \arg\max(\text{Conv}_{3\times3}(p_{\text{det}}^{0,j}))$$

- **Caption Decoder:** We employ a transformer-based decoder to describe the changes, including masked self-attention and cross-attention blocks.

$$e_w^{'} = \text{CrossAttn}(\text{SelfAttn}(e_w), p_{\text{cap}}) + e_w$$

$$\mathcal{W} = \text{softmax}(\text{FC}(e_w^{'}))$$

- **Loss Function:** We employ joint loss functions to optimize the four tasks.

$$\mathcal{L}_{\text{BCD}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}},\ \mathcal{L}_{\text{SCD}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{sim}}$$

$$\mathcal{L}_{\text{BDA}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}},\ \mathcal{L}_{\text{CC}} = \mathcal{L}_{\text{ce}}$$

Experiments

Results on Binary Change Detection Task

Method	#Params(M)	FLOPs(G)	Inference (s/sample)	LEVIR-CD			WHU-CD			CLCD		
				F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
DTCDSCN [51]	41.07 (250%)	20.44 (83%)	0.029	87.43	77.67	98.75	79.92	66.56	98.05	57.47	40.81	94.59
SNUNet [19]	12.04 (73%)	54.82 (222%)	0.036	88.16	78.83	98.82	83.22	71.26	98.44	60.82	43.63	94.90
ChangeFormer [1]	41.03 (250%)	202.79 (822%)	0.081	90.40	82.48	99.04	87.39	77.61	99.11	61.31	44.29	94.98
BIT [7]	3.55 (22%)	10.63 (43%)	0.027	89.31	80.68	98.92	83.98	72.39	98.52	59.93	42.12	94.77
ICIFNet [23]	23.82 (145%)	25.36 (103%)	0.049	89.96	81.75	98.99	88.32	79.24	98.96	68.66	52.27	95.77
Changer † [20]	11.39 (69%)	11.86 (48%)	0.025	90.70	82.99	99.08	89.18	80.48	99.17	67.07	50.46	95.69
DMINet [24]	6.24 (38%)	14.42 (58%)	0.036	90.71	82.99	99.07	88.69	79.68	98.97	67.24	50.65	95.21
GASNet † [85]	23.59 (143%)	23.52 (95%)	0.028	90.52	83.48	99.07	91.75	84.76	99.34	63.84	46.89	94.01
AMTNet [50]	16.44 (100%)	24.67 (100%)	0.032	90.76	83.08	98.96	92.27	85.64	99.32	75.10	60.12	96.45
EATDer † [56]	6.61 (40%)	23.43 (95%)	0.034	91.20	83.80	98.75	90.01	81.97	98.58	72.01	56.19	96.11
Change3D	**1.54** (9%)	**8.29** (34%)	**0.015**	**91.82**	**84.87**	**99.17**	**94.56**	**89.69**	**99.57**	**78.03**	**63.97**	**96.87**

Results on Semantic Change Detection Task

Method	#Params(M)	FLOPs(G)	Inference (s/sample)	HRSCD				SECOND			
				F1	mIoU	OA	SeK	F1	mIoU	OA	SeK
HRSCD-S4 [14]	13.71 (59%)	43.69 (23%)	0.024	69.39	67.79	81.32	22.50	58.21	71.15	86.62	18.80
ChangeMask † [89]	8.73 (37%)	37.16 (20%)	0.021	70.59	67.56	81.65	23.43	59.74	71.46	86.93	19.50
SCDNet [60]	39.62 (169%)	116.98 (62%)	0.032	70.80	67.43	81.46	23.61	60.01	70.97	87.40	19.73
SSCD-L † [15]	23.39 (100%)	189.57 (100%)	0.029	69.69	66.21	81.05	21.88	61.22	72.60	87.19	21.86
Bi-SRNet † [15]	23.39 (100%)	189.91 (100%)	0.031	71.72	67.83	82.06	24.59	61.85	72.08	87.20	21.36
MTSCD [11]	94.62 (405%)	290.28 (153%)	0.028	67.56	63.71	72.06	9.03	60.23	71.68	87.04	20.57
JFRNet [4]	23.29 (100%)	47.38 (25%)	0.035	66.65	64.65	76.20	18.78	62.63	72.82	87.10	22.56
DEFO † [44]	26.02 (111%)	401.09 (211%)	0.025	66.49	65.67	81.36	19.20	61.18	72.39	87.01	21.08
Change3D	**1.66** (7%)	**15.19** (8%)	**0.018**	**73.29**	**68.67**	**82.57**	**26.85**	**62.83**	**72.95**	**87.42**	**22.98**

Results on Building Damage Assessment Task

Method	#Params(M)	FLOPs(G)	Inference (s/sample)	F_1^{loc}	F_1^{cls}	$F_1^{overall}$	Damage F_1 Per-class			
							Non	Minor	Major	Destroy
xBD baseline [28]	25.72 (99%)	23.15 (24%)	0.021	80.47	3.42	26.54	66.31	14.35	0.94	46.57
Weber et al. [77]	48.29 (186%)	37.48 (39%)	0.029	83.60	70.02	74.10	90.60	49.30	72.20	83.70
ChangeOS-R18 [88]	15.99 (62%)	69.65 (73%)	0.019	84.62	69.87	74.30	88.61	52.10	70.36	79.65
ChangeOS-R34 [88]	26.10 (100%)	74.50 (78%)	0.021	85.16	70.28	74.74	88.63	52.38	71.16	80.08
ChangeOS-R50 [88]	53.02 (204%)	101.35 (106%)	0.025	85.41	70.88	75.24	88.98	53.33	71.24	80.60
ChangeOS-R101 [88]	72.01 (277%)	111.10 (116%)	0.031	85.69	71.14	75.50	89.11	53.11	72.44	80.79
DamFormer [8]	31.38 (121%)	220.68 (230%)	0.065	**86.86**	72.81	77.02	89.86	56.78	72.56	80.51
PCDASNet [72]	26.00 (100%)	95.9 (100%)	0.020	85.48	73.83	77.33	90.12	55.67	75.74	83.91
Change3D	**1.60** (6%)	**11.74** (12%)	**0.016**	85.74	**76.71**	**79.42**	**95.08**	**58.70**	**76.50**	**86.76**

Results on Change Captioning Task

Method	#Params(M)	FLOPs(G)	Inference (s/sample)	LEVIR-CC							DUBAI-CC						
				B-1	B-2	B-3	B-4	M	R	C	B-1	B-2	B-3	B-4	M	R	C
DUDA [58]	80.31 (201%)	20.28 (84%)	0.014	81.44	72.22	64.24	57.79	37.15	71.04	124.32	58.82	43.59	33.63	25.39	22.05	48.34	62.78
MCCFormer-S [61]	162.55 (407%)	25.09 (104%)	0.015	79.90	70.87	62.80	56.31	36.17	69.46	120.46	52.97	37.02	27.62	22.57	18.64	43.29	53.81
MCCFormer-D [61]	162.55 (407%)	25.09 (104%)	0.016	80.42	70.87	62.86	56.38	36.37	69.32	120.44	64.65	50.45	39.36	29.48	25.09	51.27	66.51
RSICCformer [47]	172.80 (433%)	27.10 (114%)	0.010	84.72	74.96	67.52	62.11	38.80	74.22	132.62	67.92	53.61	41.37	31.28	25.41	51.96	66.54
PromptCC [48]	408.58 (1024%)	19.88 (84%)	0.013	83.66	75.73	69.10	63.54	38.82	73.72	136.44	70.03	58.41	**49.44**	**40.32**	26.48	55.82	85.44
SEN [90]	39.90 (100%)	24.04 (100%)	0.008	85.10	77.05	70.01	64.09	39.59	74.57	136.02	64.12	50.41	40.28	31.01	23.67	48.19	65.15
Change3D	**5.05** (13%)	**2.39** (10%)	**0.007**	**85.81**	**77.81**	**70.57**	**64.38**	**40.03**	**75.12**	**138.29**	**72.25**	**58.68**	47.13	36.80	**27.06**	**56.04**	**86.19**

Feature Visualization

F_1 and F_2 denote the bi-temporal features, F_C represents the extracted changes.