南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Seg2Change: Adapting Open-Vocabulary Semantic Segmentation Model for Remote Sensing Change Detection

You Su
Xi'an Jiaotong University
Xi'an, Shaanxi, China
yousu@stu.xjtu.edu.cn

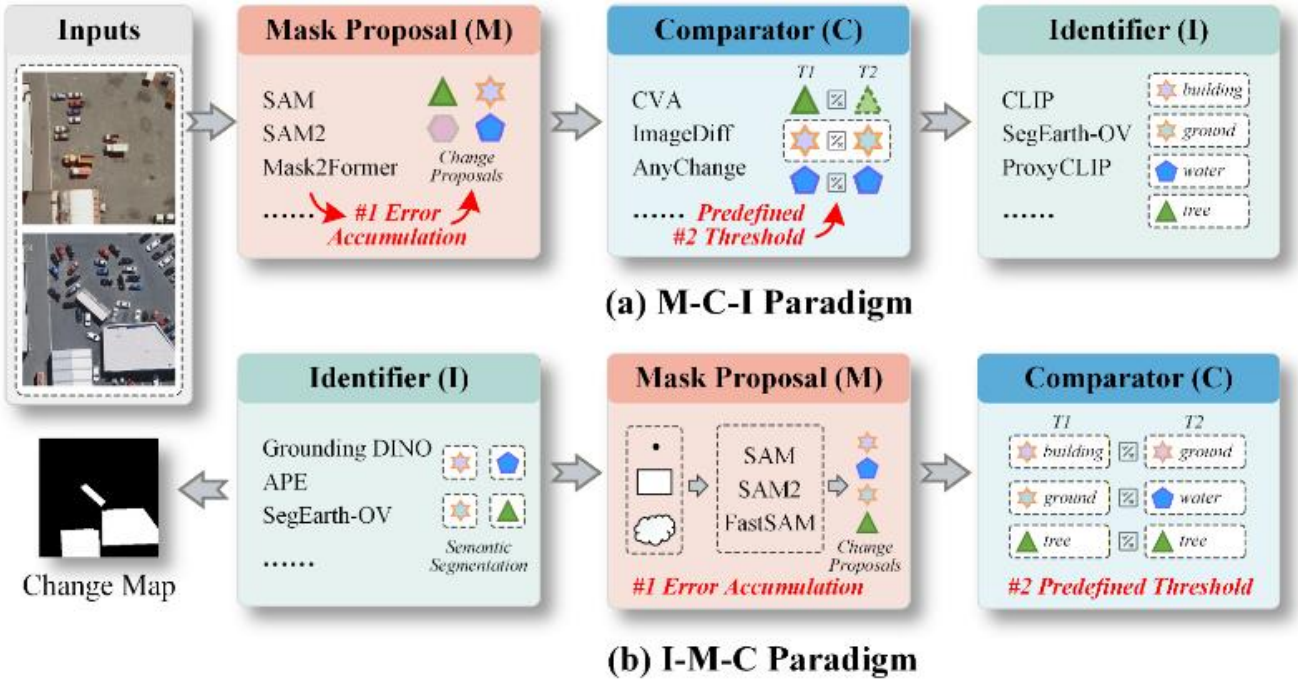
Yonghong Song
Xi'an Jiaotong University
Xi'an, Shaanxi, China
songyh@xjtu.edu.cn

Jingqi Chen
Xi'an Jiaotong University
Xi'an, Shaanxi, China
chenjingqi@stu.xjtu.edu.cn

Zehan Wen
Xi'an Jiaotong University
Xi'an, Shaanxi, China
wenzehan@stu.xjtu.edu.cn

2026年5月27日

Introduction



范式缺陷：误差链式传递

主流方法（如M-C-I、I-M-C）高度依赖SAM等预模型生成物体候选框，初始阶段的分割误差会在后续特征匹配与变化计算中逐级放大，导致“一步错步步错”，在复杂边缘与细碎地物区域极易出现漏检与误检。

阈值依赖：僵化的判定标准

所有传统方法均采用单一固定的人工阈值来二值化像素变化概率。这种“一刀切”的策略无法适应不同地物的光谱特性差异（如裸土与植被），也难以应对成像条件变化，最终导致结果图边界粗糙、伪变化噪声大量存在。

通用性差：场景的孤岛效应

现有方案多针对特定变化类型定制开发（如建筑物倒塌 vs 地表沉降），缺乏统一的特征表达与推理框架。这使得模型在跨场景迁移时需要大量重新调参与数据适配，开发与落地成本极高，无法满足实际业务的多样化需求。

变革的核心诉求

行业亟需一种能够打破误差传递、自适应场景差异且具备通用泛化能力的新范式。这正是我们提出多模态融合框架的初衷——从根本上解决现有技术的核心痛点，实现更精准、更鲁棒的遥感影像变化分析。

Dataset

Table 17: Information on the evaluation benchmark datasets and the CA-CDD coarse label source datasets. * denotes that the initially coarse, category-restricted change annotations are refined to produce category-agnostic change maps. ** indicates that all images have been preprocessed and cropped to a resolution of 512×512.

Dataset	Study Site	Number of Image Pairs	Image Size**	Resolution	Evaluation Task
WHU-CD [28]	Christchurch, New Zealand	660 (Test)	512 × 512	0.3 m	Building CD
LEVIR-CD [12]	Texas, U.S.	512 (Test)	512 × 512	0.5 m	Building CD
DSIFN [77]	Xi'an, China	48 (Test)	512 × 512	0.03–1 m	Land-cover CD
CLCD [46]	Guangdong, China	120 (Test)	512 × 512	0.5–2 m	Land-cover CD
SECOND [70]	Hangzhou, Chengdu, Shanghai, China	1694 (Test)	512 × 512	0.5–3 m	Semantic CD
SC-SCD [64]	Longwen, Zhangzhou, China	322 (Test)	512 × 512	0.5 m	Semantic CD
SECOND [70]	Hangzhou, Chengdu, Shanghai, China	2968* (Train)	512 × 512	0.5–3 m	Semantic CD
JL1-CD [49]	Multiple regions in China	1000* (Train)	512 × 512	0.5–0.75 m	Land-cover CD
CNAM-CD [80]	23 SLNAs [80]	1000* (Train)	512 × 512	0.5 m	Land-cover CD

Dataset

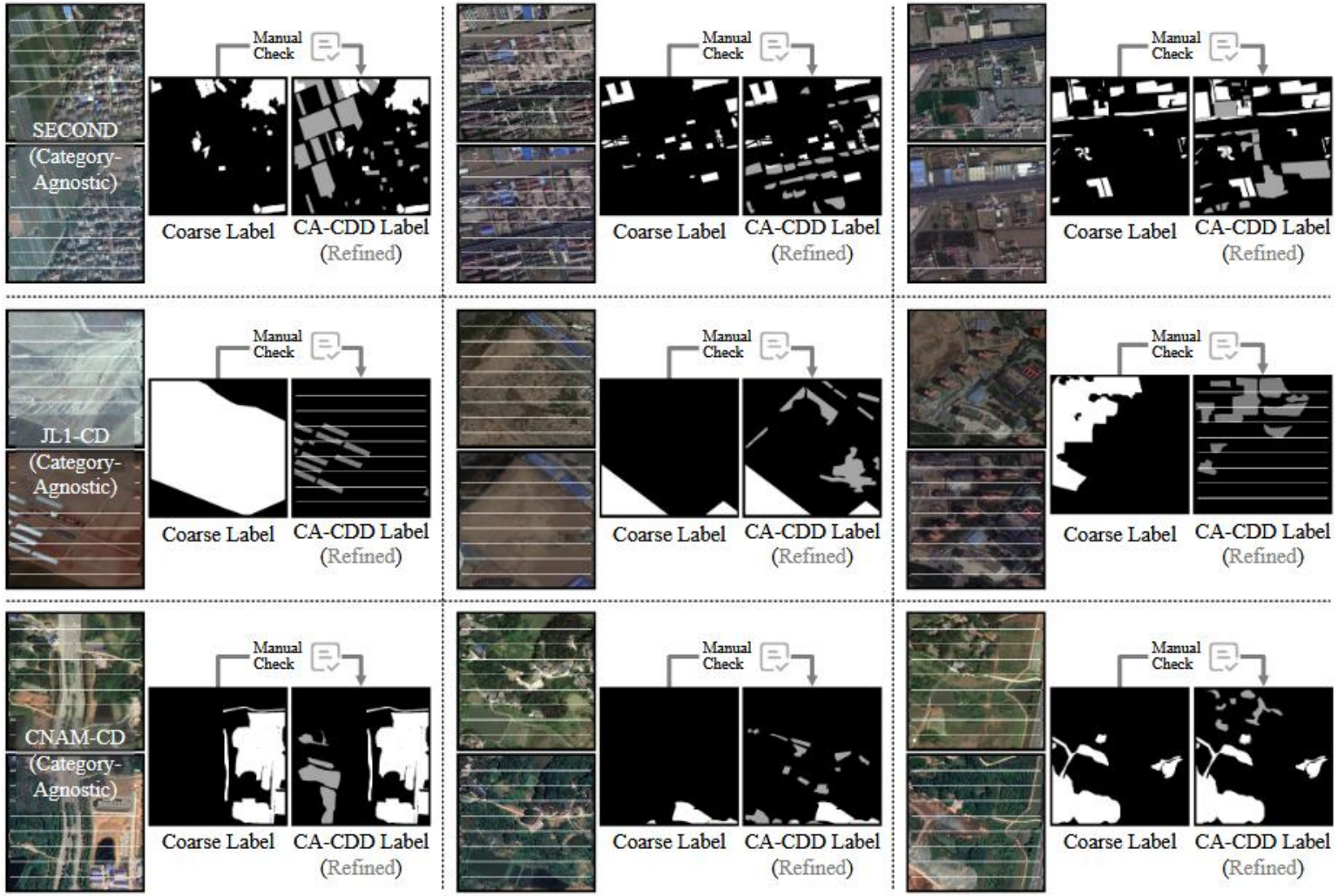


Figure 11: Visual comparison between CA-CDD category-agnostic change labels and the original coarse labels. We have improved the limited category range of the original labels. At the same time, we have refined the coarse-grained range annotations in the labels to fine-grained annotations. The gray markings in the figure represent the refinements we made to the original labels.

Methods

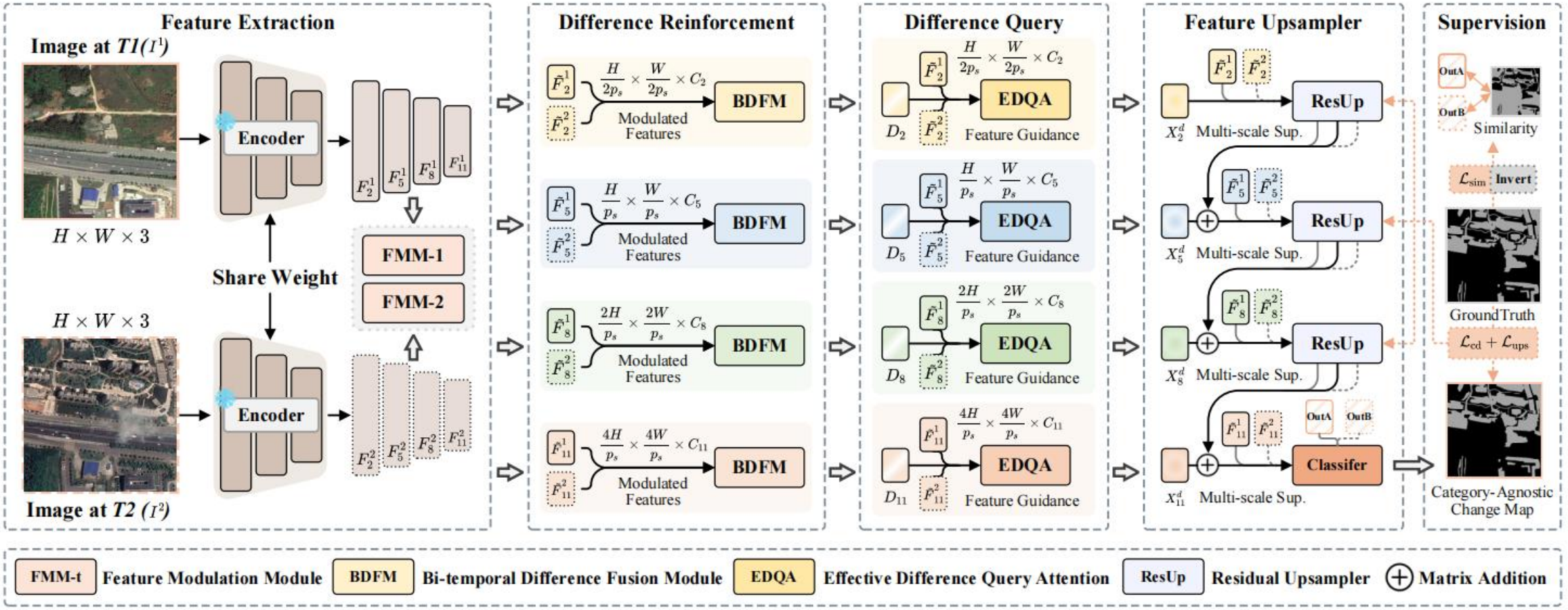
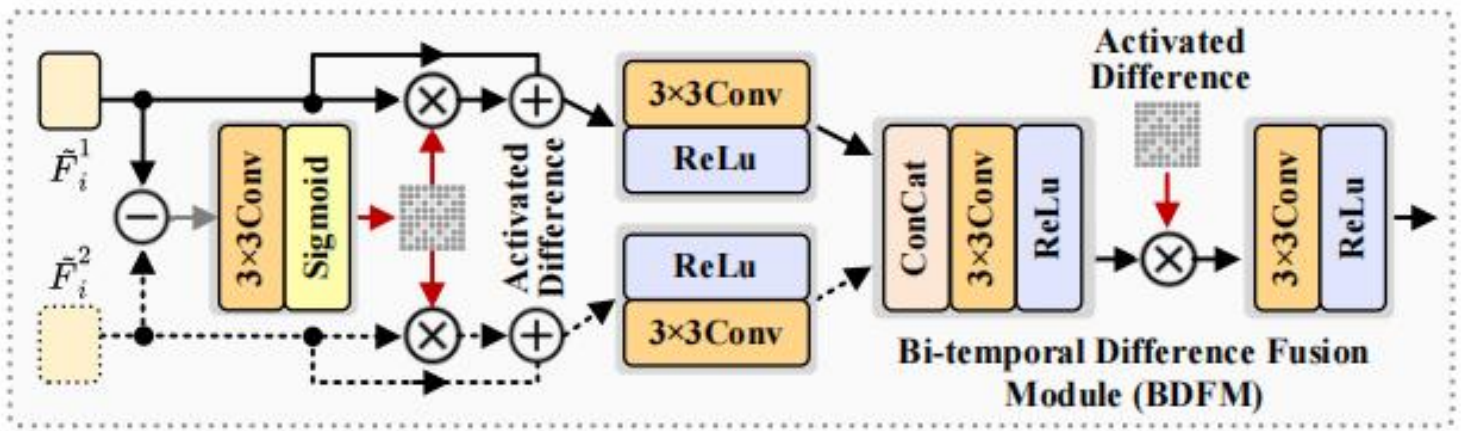


Figure 4: Category-Agnostic Change Head (CACH). Bi-temporal image features are extracted and processed through feature differences fusion and difference query modules to obtain calibrated discrepancies, which are then concatenated across multiple dimensions via our residual upsampler to produce the final category-agnostic change map.

Methods

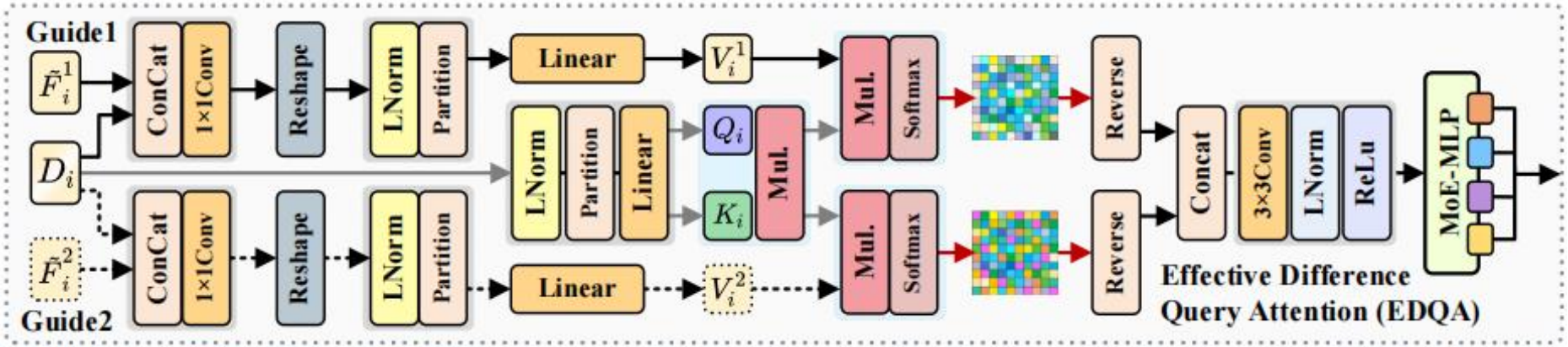


$$Att_i = \sigma(\mathrm{Conv}_{3\times3} * |\tilde{F}_i^1 - \tilde{F}_i^2|)$$

$$X_i^t = \gamma\mathrm{Conv}_{3\times3} * (\tilde{F}_i^t + Att_i \cdot \tilde{F}_i^t)$$

$$D_i = \gamma\mathrm{Conv}_{3\times3} * (\gamma\mathrm{Conv}_{3\times3} * (X_i^1 || X_i^2) \cdot Att_i)$$

Methods



$$\mathcal{L}_{\text{cd}} = \mathcal{L}_{\text{bce}}(\delta_{\uparrow}(X_2^d, X_5^d, X_8^d, X_{11}^d), y^l)$$

变化图损失：对预测变化图与标注的二元交叉熵损失

$$\mathcal{L}_{\text{ups}} = \sum_{i \in N_L} \mathcal{L}_{\text{bce}}(\delta_{\uparrow}(X_i^d), y^l)$$

上采样损失：对各层逐层预测施加多尺度监督

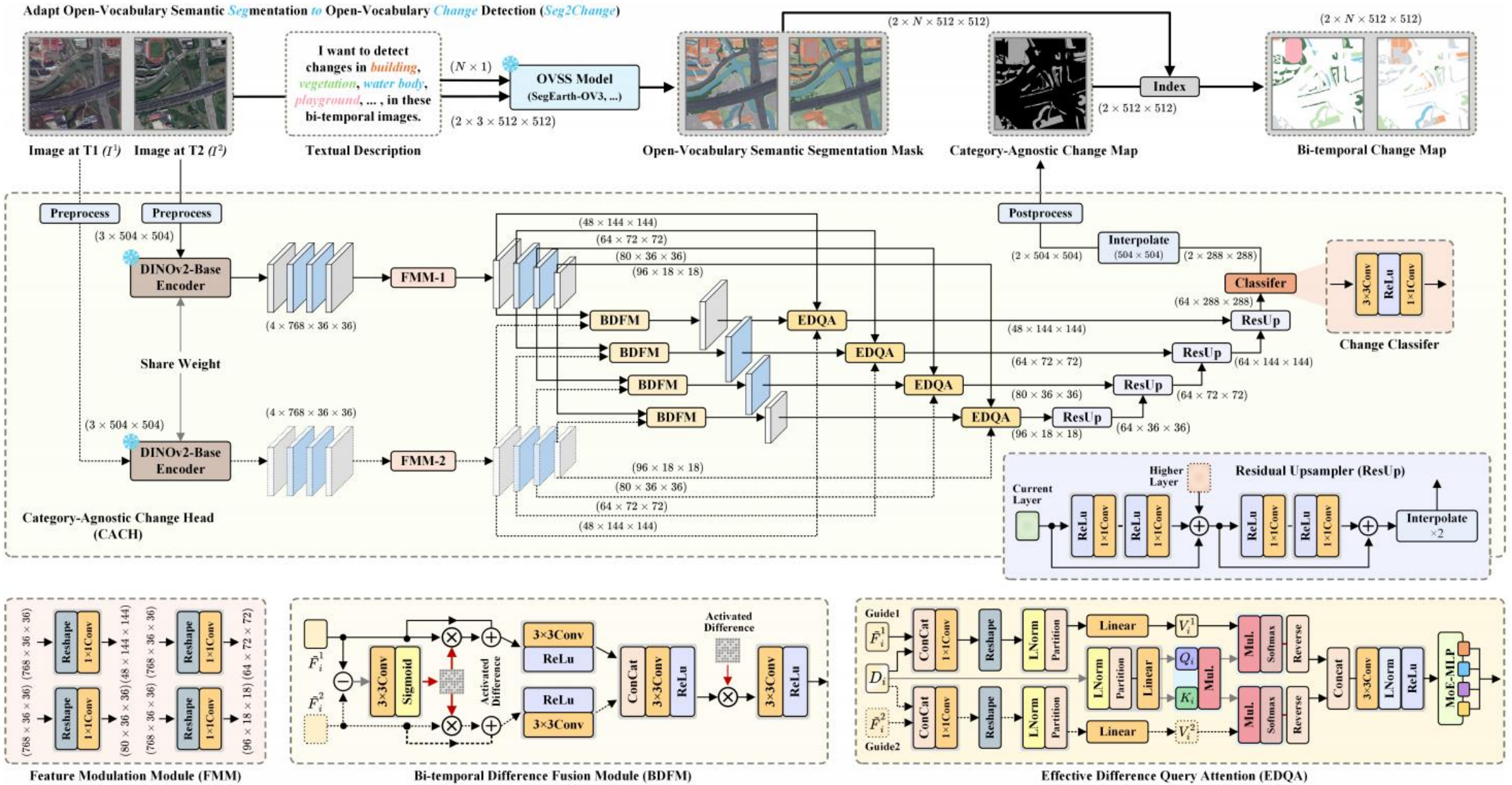
$$\mathcal{L}_{\text{sim}} = [1 - \cos(\delta_{\uparrow}(\tilde{F}^1), \delta_{\uparrow}(\tilde{F}^2))] \cdot \tilde{y}^l$$

不变区域相似度损失：对未变化区域施加余弦相似度约束，显式抑制伪变化响应

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{cd}} + \beta \mathcal{L}_{\text{ups}} + \upsilon \mathcal{L}_{\text{sim}}$$

总损失

Methods


Adapt Open-Vocabulary Semantic Segmentation to Open-Vocabulary Change Detection (Seg2Change)
I want to detect changes in building, vegetation, water body, playground, ... , in these bi-temporal images.
(N × 1)
OVSS Model (SegEarth-OV3, ...)
(2 × 3 × 512 × 512)
(2 × N × 512 × 512)
Index
(2 × 512 × 512)
(2 × N × 512 × 512)
Image at T1 (I1)
Image at T2 (I2)
Textual Description
Open-Vocabulary Semantic Segmentation Mask
Category-Agnostic Change Map
Bi-temporal Change Map
Preprocess
Preprocess
(3 × 504 × 504)
DINOv2-Base Encoder
Share Weight
DINOv2-Base Encoder
(4 × 768 × 36 × 36)
FMM-1
FMM-2
(48 × 144 × 144)
(64 × 72 × 72)
(80 × 36 × 36)
(96 × 18 × 18)
BDFM
EDQA
ResUp
Postprocess
Interpolate (504 × 504)
(2 × 288 × 288)
Classifer
(64 × 288 × 288)
Change Classifer
Category-Agnostic Change Head (CACH)
Residual Upsampler (ResUp)
Current Layer
Higher Layer
ReLu
1×1Conv
Interpolate ×2
Reshape
Sigmoid
3×3Conv
ConCat
Activated Difference
Guide1
Guide2
LNorm
Partition
Linear
Mul.
Softmax
Reverse
Concat
MoE-MLP
Feature Modulation Module (FMM)
Bi-temporal Difference Fusion Module (BDFM)
Effective Difference Query Attention (EDQA)

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 1: Performance comparison on binary OVCD datasets. † and ‡ denote the M–C–I and I–M–C configurations of DynamicEarth. * denotes the variants implemented with SegEarth-OV3 (SegE-OV3 for short). The best and the second-best results are, respectively, marked in BOLD and in underline. Except for Kappa, all results are expressed in percentage (%).

Method	Identifier	Comparator	WHU-CD			LEVIR-CD			DSIFN			CLCD		
			$F1^c$	IoU^c	Kappa	$F1^c$	IoU^c	Kappa	$F1^c$	IoU^c	Kappa	$F1^c$	IoU^c	Kappa
PCA_KMeans [9]	/	KMeans [9]	14.33	7.72	0.0846	9.96	5.24	0.0116	35.48	21.57	0.1865	18.42	10.14	0.0780
CVA [6]	/	CVA Match [6]	7.17	3.72	0.0023	9.27	4.86	0.0017	27.88	16.20	0.0142	13.39	7.17	0.0017
DCVA [3]	/	CVA Match [6]	20.55	11.45	0.1537	13.85	7.44	0.0507	38.52	23.85	0.1758	20.81	11.62	0.0918
UCD-SCM [63]	SAM	OTSU [53]	32.13	19.14	0.2792	32.36	19.30	0.2734	40.13	25.10	0.2443	23.31	13.19	0.1522
AnyChange [79]	SAM	Latent Match [79]	28.13	16.37	0.2330	32.68	19.53	0.2672	39.19	24.37	0.2280	31.96	19.02	0.2342
AnyChange* [79]	SegE-OV3	Latent Match [79]	69.25	52.96	0.6790	72.27	56.58	0.7043	54.69	37.64	0.4798	27.55	15.98	0.2276
Inst-CEG [39]	APE	CEG [39]	62.54	45.49	0.6074	63.29	46.30	0.6084	31.81	18.91	0.2430	6.76	3.50	0.0020
Inst-CEG* [39]	SegE-OV3	CEG [39]	71.35	55.46	0.7016	70.62	54.58	0.6912	47.21	30.90	0.3985	10.09	5.32	0.0648
DynamicEarth† [38]	SAM2	DINOv2 [52]	57.35	40.20	0.5541	46.43	30.23	0.4242	54.35	37.32	0.4741	23.83	13.52	0.1916
DynamicEarth‡ [38]	APE	DINOv2 [52]	75.85	61.09	0.7488	69.70	53.50	0.6789	26.42	15.22	0.2260	14.97	8.09	0.1382
DynamicEarth* [38]	SegE-OV3	DINOv2 [52]	79.66	66.20	0.7879	71.97	56.21	0.7031	39.30	24.45	0.3423	38.16	23.58	0.3535
Seg2Change	SegE-OV3	CACH (ours)	**86.18**	**75.72**	**0.8562**	**78.72**	**64.91**	**0.7742**	**58.56**	**41.40**	**0.5075**	**47.89**	**31.48**	**0.4239**

Experiments

Figure 6: Open-vocabulary building and land-cover change detection examples. In each group: images at T1 (I^1), T2 (I^2), labels, the results of DynamicEarth and our Seg2Change. Color rendering: *"Building/Land-cover"*.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 2: Performance, computational resource usage, and inference time cost comparison on semantic OVCD datasets. GPU memory usage and inference time are evaluated on a pair of bi-temporal remote sensing images ([2, 3, 512, 512]).

Method	GPU Memory Usage (GB) ↓	Inference (ms/sample) ↓	SC-SCD				SECOND			
			$mF1^c$	$mIoU^c$	mOA	mKappa	$mF1^c$	$mIoU^c$	mOA	mKappa
UCD-SCM [63]	9.46 (100%)	3225 (100%)	12.06	6.51	85.95	0.0619	14.63	8.40	78.61	0.0795
AnyChange [79]	6.59 (69%)	3988 (124%)	16.51	9.32	78.27	0.1056	19.54	11.84	76.07	0.1252
AnyChange* [79]	10.15 (107%)	2683 (83%)	16.64	9.48	88.06	0.1192	21.30	13.81	77.97	0.1492
Inst-CEG [39]	14.95 (158%)	5672 (176%)	6.90	3.68	94.40	0.0551	17.82	10.46	92.80	0.1604
Inst-CEG* [39]	7.08 (75%)	2928 (91%)	23.57	14.35	94.49	0.2139	29.35	18.40	93.62	0.2670
DynamicEarth† [38]	7.33 (77%)	5035 (156%)	29.11	17.97	91.37	0.2532	37.51	23.58	91.88	0.3297
DynamicEarth‡ [38]	15.33 (162%)	6784 (210%)	9.19	5.16	95.25	0.0877	22.17	13.72	93.76	0.2109
DynamicEarth* [38]	11.19 (118%)	2892 (89%)	19.87	11.47	95.03	0.1804	25.43	15.24	93.83	0.2286
Seg2Change	**6.08** (64%)	**1521** (47%)	**35.68**	**23.22**	**95.82**	**0.3385**	**42.89**	**29.08**	**95.17**	**0.4045**

Experiments

Figure 7: Open-vocabulary semantic change detection examples. In each group: images at T1 (I^1), T2 (I^2), labels, the results of DynamicEarth and our Seg2Change. Color rendering: *"Building"*, *"Vegetation"*, *"Tree"*, *"Playground"*, *"Water"*, *"Bareland"*.

Experiments

Table 3: Investigation on the impact of the proposed CACH components on the WHU-CD and DSIFN datasets.

Components in CACH				WHU-CD			DSIFN		
The Effectiveness of Difference Modules in CACH									
FMM	BDFM	EDQA	ResUp	$F1^c$	IoU^c	Kappa	$F1^c$	IoU^c	Kappa
✓	✗	✗	✗	69.68	53.47	0.6870	49.65	33.02	0.3581
✓	✓	✗	✗	78.10	64.06	0.7732	53.22	36.26	0.4758
✓	✓	✓	✗	84.18	72.68	0.8354	55.57	38.47	0.4966
✓	✓	✓	✓	**86.18**	**75.72**	**0.8562**	**58.56**	**41.40**	**0.5075**
The Effectiveness of Backbone Features in CACH									
Shallow: $\{F_0, F_2, F_3, F_5\}$				78.13	64.12	0.7736	49.78	33.14	0.4780
Deep: $\{F_6, F_8, F_9, F_{11}\}$				81.41	68.64	0.8069	54.35	37.32	0.5318
Comb1: $\{F_1, F_4, F_6, F_{10}\}$				82.92	70.82	0.8224	55.93	38.82	0.5390
Comb2: $\{F_2, F_5, F_8, F_{11}\}$				**86.18**	**75.72**	**0.8562**	**58.56**	**41.40**	**0.5075**
The Effectiveness of Loss Functions in CACH									
Change Map Loss ($\mathcal{L}_{cd}$)				84.31	72.88	0.8371	52.69	35.77	0.7593
+ Upsample Loss ($\mathcal{L}_{ups}$)				85.58	74.79	0.8500	55.32	38.24	0.4942
+ Unchanged Loss ($\mathcal{L}_{sim}$)				**86.18**	**75.72**	**0.8562**	**58.56**	**41.40**	**0.5075**

Thanks