BiRNeT: A Unified Bidirectional Reasoning Paradigm for Open-Vocabulary Remote Sensing Change Detection

Anonymous authors
Paper under double-blind review

Abstract

Open-vocabulary change detection (OVCD) aims to identify changes of arbitrary categories based on text descriptions, overcoming the reliance of conventional methods on predefined categories. However, existing approaches generally treat change localization and semantic recognition as separate processes, and the lack of interaction between them has become a key bottleneck for further performance improvement. In this paper, we propose BiRNet, a unified bidirectional reasoning paradigm. It uses change probabilities to constrain bi-temporal semantic information and performs joint inference to achieve unified prediction of change states and change categories. To support this paradigm, an Adaptive Multi-scale Context-Difference Fusion Module (AMCDF) is designed to enhance change representation. A Bidirectional Change-Constrained Semantic Transition Module (BCCST) is further proposed to couple localization and recognition into a joint decision through a semantic transition matrix constrained by change probabilities. BCCST requires no additional training and offers plug-and-play capability. Experimental results on four benchmark datasets demonstrate that BiRNet achieves competitive performance, with significant $F1^C$ gains of 23.69% and 6.93% on S2Looking and SECOND, respectively. The code will be released upon publication.

1 Introduction

Remote sensing change detection (RSCD) is an important technique for observing and analyzing surface dynamics, with broad applications in urban expansion monitoring (Zhang et al., 2025; Ning et al., 2024), disaster assessment (Gao et al., 2025), environment monitoring (Friedlander et al., 2026), and land-cover change (Chen & Shi, 2020). Existing methods are mainly based on supervised learning (Zhao et al., 2024), which typically requires large-scale pixel-level annotations. However, their recognition capabilities are constrained by predefined categories, making them difficult to adapt to open-vocabulary scenarios (Zhu & Yang, 2025).

In recent years, the advancement of vision-language models (VLMs) has provided new opportunities to overcome these limitations (Li et al., 2024; Tang et al., 2026; Kuang et al., 2026), driving the development of open-vocabulary change detection (OVCD). OVCD aims to localize and identify changes of arbitrary categories in bi-temporal imagery based on textual descriptions (Li et al., 2026).

Seg2Change (Su et al., 2026) is a representative paradigm for OVCD, as illustrated in Figure 1(a). It employs a category-agnostic change head to generate a change map and combines it with bi-temporal semantic predictions from an open-vocabulary semantic segmentation (OVSS) model to identify change categories. This paradigm effectively exploits the open-vocabulary capability of OVSS (Peng et al., 2026). However, change localization and semantic recognition are processed independently and lack effective interaction, which constrains further performance improvement.

In fact, human judgments of scene changes rely not only on visual differences but also on object semantic information (Spotorno et al., 2013; Nentwich et al., 2023). Motivated by these findings, this paper proposes BiRNet, a unified bidirectional reasoning paradigm for open-vocabulary remote sensing change detection. As illustrated in Figure 1(b), BiRNet comprises a change detection branch and a semantic segmentation branch, and jointly utilizes binary change probabilities and bi-temporal semantic information to determine the final semantic change results.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

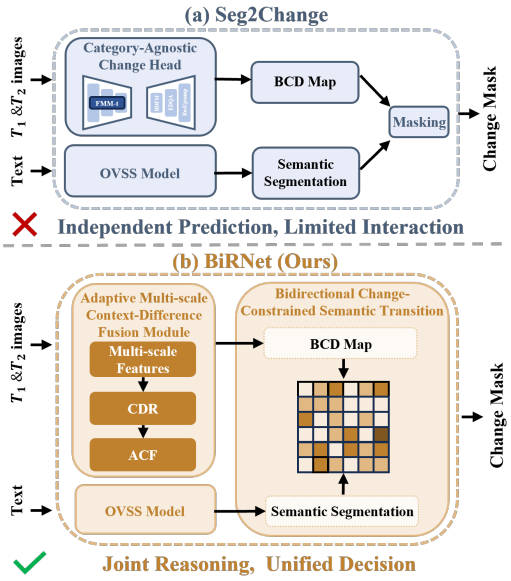

Figure 1: Paradigm comparison between Seg2Change and BiRNet.

Specifically, the change detection branch employs an Adaptive Multi-scale Context-Difference Fusion Module (AMCDF) to integrate high-level contextual and difference information with low-level difference features. Through adaptive fusion based on local and global information, AMCDF generates binary change probabilities. Meanwhile, the semantic segmentation branch obtains bi-temporal semantic information through an OVSS model. Subsequently, the Bidirectional Change-Constrained Semantic Transition Module (BCCST) constructs pixel-level semantic transition matrices and constrains them using change probabilities. The most probable semantic transition is then selected to simultaneously determine the change state and the semantic categories before and after the change.

Unlike existing methods that separately perform change localization and category recognition, BiRNet allows change probabilities to constrain semantic reasoning while using bi-temporal semantic information to assist change-state determination. This enables unified prediction of change locations and semantic categories before and after the change while preserving the open-vocabulary capability of OVSS. The main contributions are summarized as follows:

1. BiRNet is proposed to jointly infer change states and semantic categories by integrating binary change probabilities with bi-temporal semantic information, while supporting flexible adaptation to different BCD and OVSS models.
2. We introduce AMCDF, which adaptively fuses high-level contextual and difference information with low-level features to enhance change representation and reduce false alarms.
3. We introduce BCCST to jointly leverage change localization and semantic information for semantic change prediction. It improves prediction reliability without additional training and supports plug-and-play deployment.

2 Related Work

Vision Foundation Models. In recent years, vision foundation models (VFMs) have learned general visual representations through large-scale pretraining, providing strong feature foundations for downstream vision tasks (Liu et al., 2024; Lu et al., 2025; Shi & Yang, 2024). SAM (Kirillov et al., 2023) performs object segmentation through prompts such as points and bounding boxes, while DINO (Caron et al., 2021), DINOv2 (Oquab et al., 2023), and DINOv3 (Siméoni et al., 2025) learn visual representations with strong semantic and spatial structures through self-distillation. Notably, DINOv3 demonstrates strong feature extraction capabilities on various remote sensing segmentation tasks, even without dedicated pretraining on remote sensing imagery. Building upon these advances, vision-language models (VLMs) further establish connections between visual features and textual semantics, providing a foundation for open-vocabulary visual understanding. CLIP (Radford et al., 2021) supports open-vocabulary recognition through contrastive learning on large-scale image-text pairs, while DINO.txt (Jose et al., 2025) aligns the visual representations of DINOv2 with textual semantics, preserving its visual representation capability while incorporating open-vocabulary semantic information.

Open-Vocabulary Semantic Segmentation. Open-vocabulary semantic segmentation (OVSS) aims to recognize and segment arbitrary categories based on text descriptions (Wang et al., 2025; Park et al., 2025). Early methods mainly leverage the image-text alignment of CLIP for pixel-level prediction of unseen categories (Kang et al., 2025; Wu et al., 2024). Subsequent approaches incorporate VFMs such as DINO and SAM to enhance spatial representations (Lan et al., 2024) and further

develop training-free open-vocabulary segmentation methods (Shi et al., 2025). In recent years, OVSS for remote sensing imagery has attracted increasing attention. However, the differences between remote sensing and natural images in terms of object scale, shape, and scene distribution make direct transfer of natural image models prone to inaccurate localization and unclear boundaries (Cao et al., 2025; Huang et al., 2025). To address this issue, SegEarth-OV (Li et al., 2025a) improves performance by enhancing the spatial representation of CLIP features, SegEarth-OV3 (Li et al., 2025b) exploits the multi-head outputs of SAM 3 (Carion et al., 2026) to enhance open-vocabulary segmentation, and CAFe-DINO (Faulkenberry & Prasad, 2026) combines the visual representations of DINOv3 with the vision-language alignment of DINO.txt to provide stable bi-temporal semantic predictions.

Open-Vocabulary Change Detection. Existing OVCD methods mainly follow three paradigms: M–C–I (Li et al., 2026), I–M–C (Li et al., 2026), and Seg2Change (Su et al., 2026). Representative M–C–I methods include OpenDPR (Guo et al., 2026), which enhances category recognition through visual prototype retrieval, and AdaptOVCD (Dou et al., 2026), which improves detection stability through multi-level adaptive information fusion. The representative I–M–C method, OmniOVCD (Zhang et al., 2026), exploits the concept segmentation capability of SAM 3 to generate object-level semantic representations and performs change detection through instance-level matching. Both paradigms rely on a mask generator to produce change proposals, making them susceptible to mask errors and limiting their applicability across different change detection tasks (Su et al., 2026; Peng et al., 2026). Seg2Change generates a change map using a category-agnostic change head (CACH) and combines it with bi-temporal OVSS predictions to determine change categories. However, existing methods still treat change localization and semantic recognition as relatively independent processes, with limited explicit information interaction between them. Therefore, establishing an effective coordination mechanism between change localization and semantic recognition remains an important research problem.

3 METHOD

3.1 OVERVIEW

BiRNet establishes a dual-branch architecture with a change detection branch and an OVSS branch, whose outputs are further integrated by the Bidirectional Change-Constrained Semantic Transition Module (BCCST) for unified semantic change prediction, as illustrated in Figure 2. Given bi-temporal images I_1 and I_2, the change detection branch extracts multi-scale features using a pretrained SAM ViT-B Image Encoder (Kirillov et al., 2023) and generates pixel-level change probabilities through the Adaptive Multi-scale Context-Difference Fusion Module (AMCDF). Meanwhile, the OVSS branch obtains semantic predictions for the two temporal phases from the images and user-specified category text. BCCST constructs a semantic transition matrix from these predictions and constrains it with change probabilities, thereby determining the change status and semantic categories before and after the change.

3.2 SAM-BASED MULTI-SCALE FEATURE EXTRACTION

Given a pair of bi-temporal images I_1 and I_2, we employ a pretrained SAM ViT-B Image Encoder to extract multi-scale visual features. Specifically, features from different Transformer layers are combined to exploit complementary information across depths. Features from layers 2 and 5 form L_t, while those from layers 8 and 11 form H_t:

$$L_t = S_2^t \oplus S_5^t, \qquad H_t = S_8^t \oplus S_{11}^t. \tag{1}$$

where $t \in 1, 2$ denotes the two temporal phases, and $\oplus$ denotes element-wise feature addition. These representations are further projected through 3×3 convolutions to obtain the low-level representation F_l^t and high-level representation F_h^t, respectively. Specifically, $F_l^t \in \mathbb{R}^{64\times128\times128}$ preserves local structures and spatial details for precise boundary localization, while $F_h^t \in \mathbb{R}^{128\times64\times64}$ captures richer semantic information and contextual dependencies for robust change modeling.

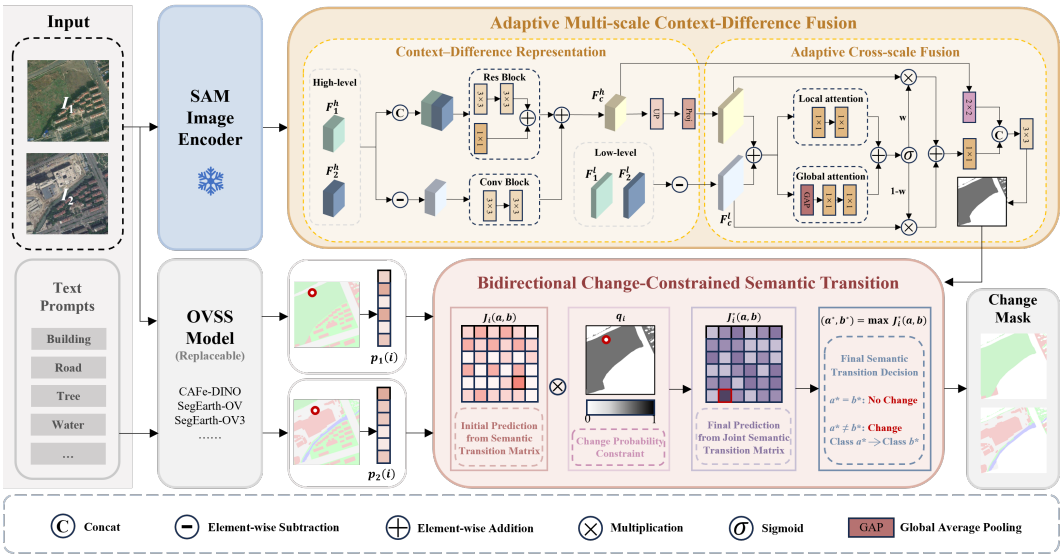

Figure 2: Overall framework of the proposed BiRNet. The proposed method consists of a change detection branch and an OVSS branch. AMCDF enhances change representation and generates binary change probabilities, while the OVSS branch provides bi-temporal semantic predictions. BCCST then constrains semantic transitions with change probabilities to jointly determine change states and semantic categories.

3.3 Adaptive Multi-scale Context-Difference Fusion

An Adaptive Multi-scale Context-Difference Fusion Module (AMCDF) is proposed to fully exploit the complementary information from high-level and low-level features. AMCDF consists of Context-Difference Representation and Adaptive Cross-scale Fusion. Through these components, more reliable change representations can be obtained.

3.3.1 Context-Difference Representation

For the bi-temporal high-level features F_h^1 and F_h^2, the element-wise absolute difference is first calculated to obtain the initial change representation, which is further refined as:

$$D_h = \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{3\times3}(|F_h^1 - F_h^2|)). \tag{2}$$

To compensate for the lack of contextual information in the difference features, the bi-temporal high-level features are concatenated along the channel dimension and processed by a residual block for contextual modeling:

$$C_h = \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{3\times3}(\mathrm{Concat}(F_h^1, F_h^2)) + \mathrm{Conv}_{1\times1}(\mathrm{Concat}(F_h^1, F_h^2)). \tag{3}$$

The difference representation D_h and contextual representation C_h are fused by element-wise addition to obtain the high-level change feature F_h^c, where the contextual representation provides bi-temporal scene information as a semantic reference, while the difference representation highlights changes between the two temporal phases.

For the low-level features, the element-wise absolute difference is directly adopted to preserve fine-grained spatial change information:

$$F_l^c = |F_l^1 - F_l^2|. \tag{4}$$

3.3.2 Adaptive Cross-scale Fusion

The high-level change feature provides stronger semantic discrimination, whereas the low-level change feature preserves finer spatial structures. To fully exploit their complementary information, the high-level change feature is first upsampled and projected along the channel dimension:

$$\widehat{F}_h^c = \mathrm{Conv}_{1\times1}(\mathrm{Up}(F_h^c)). \tag{5}$$

The low-level change feature F_l^c and the projected high-level change feature $\widehat{F}_h^c$ are integrated by element-wise addition to obtain Z. Fusion weights are subsequently generated from both local and global perspectives based on Z:

$$\begin{aligned} A_l &= \mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{1\times1}(Z)\right), \\ A_g &= \mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(Z))\right), \\ w &= \sigma(A_l + A_g). \end{aligned} \tag{6}$$

Here, $\sigma(\cdot)$ denotes the Sigmoid function, and GAP denotes global average pooling. Finally, the high-level and low-level change features are adaptively fused according to their feature responses to obtain the cross-scale change feature:

$$F_f = w \odot F_l^c + (1 - w) \odot \widehat{F}_h^c, \tag{7}$$

where $\odot$ denotes element-wise multiplication. This process dynamically adjusts the contributions of high-level and low-level change information according to their feature reliability. It preserves the spatial details of low-level features while exploiting the stable semantic information provided by high-level features, resulting in a more reliable change representation.

After obtaining F_f, the high-level change feature F_h^c is further incorporated for decoding to generate the pixel-level binary change probability q_i, where i represents a pixel in the image.

3.4 Bidirectional Change-Constrained Semantic Transition

BCCST is the core module that enables joint reasoning between change localization and semantic recognition. For an arbitrary pixel i, given the change probability q_i produced by the binary change detection branch and the bi-temporal semantic probabilities $p_1^i, p_2^i \in \mathbb{R}^K$ obtained from the OVSS branch, where K denotes the number of user-specified categories, BCCST first models the semantic transition between the two temporal phases and then constrains the transition using the change probability. The final change status and semantic categories are determined through the semantic transition reasoning process, where change information and semantic information jointly contribute to the decision. The overall process is illustrated in Figure 3.

First, a pixel-level semantic transition matrix $J_i = (p_1^i)^\mathsf{T} p_2^i \in \mathbb{R}^{K\times K}$ is constructed from the bi-temporal semantic probabilities, where $J_i(a, b)$ denotes the probability that pixel i transitions from category a at the first temporal phase to category b at the second temporal phase. The diagonal elements represent transitions between the same semantic categories across the two temporal phases and provide homogeneous evidence. In contrast, the off-diagonal elements represent transitions between different semantic categories and provide heterogeneous evidence. Accordingly, the probabilities of same-category and different-category transitions are aggregated to obtain the homogeneous evidence E_{same} and heterogeneous evidence E_{diff}, respectively:

$$E_{\text{same}} = \sum_{a=1}^{K} J_i(a, a), \qquad E_{\text{diff}} = 1 - E_{\text{same}}. \tag{8}$$

The change probability q_i is then used to constrain the semantic transition matrix, yielding the joint semantic transition matrix:

$$J_i^*(a, b) = \begin{cases} \dfrac{(1 - q_i)J_i(a, b)}{E_{\text{same}} + \epsilon}, & a = b, \\ \dfrac{q_i J_i(a, b)}{E_{\text{diff}} + \epsilon}, & a \neq b. \end{cases} \tag{9}$$

Here, ϵ is a small positive constant for numerical stability. Finally, BCCST selects the most probable category transition from the joint semantic transition matrix:

$$(a^*, b^*) = \arg\max_{a,b} J_i^*(a, b). \tag{10}$$

The selected pair (a^*, b^*) therefore represents the most probable transition of pixel i from category a^* to category b^*. When $a^* = b^*$, the pixel is considered unchanged. When $a^* \neq b^*$, the pixel is identified as changed, with the semantic transition from a^* to b^* representing the semantic categories before and after the change.

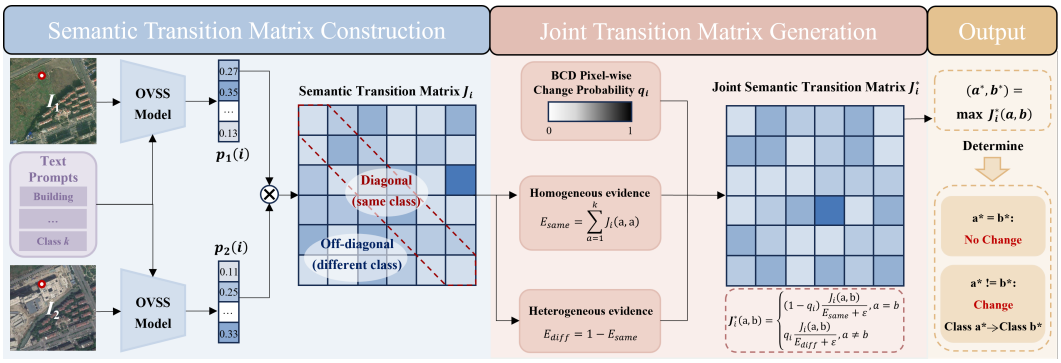

Figure 3: Bidirectional Change-Constrained Semantic Transition module (BCCST). Bi-temporal semantic predictions are modeled as a semantic transition matrix and constrained by change probabilities to jointly determine change states and semantic categories.

Therefore, BCCST simultaneously determines the change status and the semantic categories before and after the change by selecting the most reliable result from the joint semantic transition matrix. This design avoids treating change localization and category recognition as two independent stages. The change probability constrains semantic transitions, while the bi-temporal semantic relationship also contributes to the final determination of the change status. In this way, BCCST preserves the open-vocabulary recognition capability of OVSS while enabling unified prediction of change locations and semantic categories across the two temporal phases.

3.5 Loss Function

Only the change detection branch is trained, while the pretrained SAM Image Encoder remains frozen throughout the training process. To supervise the pixel-level prediction of change probabilities and alleviate the class imbalance between changed and unchanged regions, binary cross-entropy loss and Dice loss are employed:

$$\begin{aligned} \mathcal{L}_{\text{BCE}} &= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log q_i + (1-y_i)\log(1-q_i)\right], \\ \mathcal{L}_{\text{Dice}} &= 1 - \frac{2\sum_{i=1}^{N} q_i y_i + 1}{\sum_{i=1}^{N} q_i + \sum_{i=1}^{N} y_i + 1}. \end{aligned} \tag{11}$$

Here, N denotes the total number of pixels involved in the loss computation. The final training loss for the binary change detection branch is defined as

$$\mathcal{L}_{\text{BCD}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}. \tag{12}$$

4 Experiments

4.1 Datasets

To comprehensively evaluate BiRNet, experiments are conducted on four diverse remote sensing change detection datasets: LEVIR-CD (Chen & Shi, 2020), WHU-CD (Ji et al., 2019), S2Looking (Shen et al., 2021), and SECOND (Yang et al., 2022). LEVIR-CD contains 637 bi-temporal image pairs with a spatial resolution of 0.5 m/pixel and an image size of 1024×1024, where 128 pairs are used for testing. WHU-CD consists of 1,950 high-resolution aerial image pairs at a spatial resolution of 0.2 m/pixel, cropped into 512×512 patches, with 690 pairs used for testing. S2Looking provides 5,000 bi-temporal image pairs with spatial resolutions ranging from 0.5 to 0.8 m/pixel, including 1,000 pairs in the test set. SECOND contains 4,662 bi-temporal aerial image pairs with pixel-wise annotations for six land-cover categories, including surface, tree, low vegetation, water, building, and playground, where 1,694 pairs are used for testing.

Table 1: Performance comparison on building OVCD datasets. The best and second-best results are marked in bold and underline, respectively.

Method	LEVIR-CD (%)				WHU-CD (%)				S2Looking (%)			
	Prec.	Rec.	IoU^c	F1^c	Prec.	Rec.	IoU^c	F1^c	Prec.	Rec.	IoU^c	F1^c
SAM-DINOv2-SegOV	45.98	65.01	36.86	53.86	64.94	70.17	50.89	67.45	39.92	39.05	24.59	39.48
SAM2-DINOv2-SegOV	43.18	60.01	33.53	50.22	71.53	41.57	35.67	52.58	38.68	36.99	23.32	37.82
APE-/-DINO	62.94	78.09	53.50	69.70	86.77	67.90	61.53	76.18	45.39	11.48	10.09	18.32
APE-/-DINOv2	56.50	81.30	50.00	66.67	88.13	76.30	69.19	81.79	6.98	18.12	5.31	10.08
OpenDPR	52.83	74.74	44.82	61.90	68.66	70.81	53.52	69.72	29.18	55.85	23.71	38.34
AdaptOVCD	62.83	74.10	51.52	68.00	80.60	72.85	61.99	76.53	42.97	36.19	24.45	39.29
OmniOVCD	73.97	87.94	67.16	80.35	82.52	90.73	76.11	86.43	32.57	48.93	24.31	39.11
ReA-OVCD	74.01	**93.29**	70.26	82.54	88.44	**92.22**	**82.30**	**90.29**	29.37	**60.62**	24.67	39.57
Seg2Change	69.71	90.38	64.89	78.71	80.74	92.02	75.46	86.02	56.05	25.21	21.05	34.77
Ours	**86.85**	85.69	**75.85**	**86.27**	**92.68**	85.93	80.46	89.17	**72.26**	56.25	**46.26**	**63.26**

Table 2: Performance comparison on six land-cover classes of the SECOND dataset. The best and second-best results are marked in bold and underline, respectively. "–" denotes a score close to zero.

Method	Building		Tree		Water		Low vegetation		Surface		Playground		Class Avg.	
	IoU^c	F1^c	IoU^c	F1^c	IoU^c	F1^c	IoU^c	F1^c	IoU^c	F1^c	IoU^c	F1^c	mIoU	mF1
SAM-DINOv2-SegOV	38.54	55.63	15.68	27.10	15.34	26.61	21.40	35.26	27.86	43.57	23.14	37.59	23.66	37.63
SAM2-DINOv2-SegOV	36.66	53.65	16.47	28.29	14.66	25.57	20.02	33.36	19.56	32.72	25.11	40.15	22.08	35.62
APE-/-DINO	29.08	45.06	9.93	18.06	12.34	21.97	–	–	–	–	25.59	40.75	12.82	20.97
APE-/-DINOv2	31.74	48.19	10.89	19.64	12.34	21.97	–	–	–	–	26.93	42.43	13.65	22.04
OpenDPR	44.03	61.14	16.86	28.85	26.18	41.50	22.77	37.09	33.17	49.81	39.59	56.72	30.43	45.85
AdaptOVCD	44.98	62.05	12.71	22.56	23.36	37.87	22.11	36.21	34.03	50.78	25.58	40.74	27.13	41.70
OmniOVCD	45.20	62.26	16.68	28.59	21.20	34.98	24.48	39.33	27.71	43.40	26.94	42.44	27.04	41.83
ReA-OVCD	56.18	71.94	19.35	32.42	26.22	41.55	**27.99**	**43.74**	34.81	51.65	24.41	39.25	31.49	46.76
Seg2Change	**61.77**	**76.37**	16.04	27.65	23.10	37.54	18.17	30.75	37.26	54.29	17.72	30.11	29.01	42.79
Ours	61.00	75.77	**20.78**	**34.41**	**27.46**	**43.09**	24.23	39.01	**45.16**	**62.22**	**51.10**	**67.64**	**38.29**	**53.69**

4.2 Experimental Setup

All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. The change detection branch is trained with the Adam optimizer for 50 epochs using an initial learning rate of 1×10^{-3}. The OVSS branch adopts CAFe-DINO with pretrained weights directly loaded, without additional training or fine-tuning. We evaluate performance using Precision, Recall, $\mathrm{F1}^C$, and IoU^C, calculated for each category.

4.3 Comparison with State-of-the-Art Methods

We compare BiRNet with state-of-the-art OVCD methods, including the M–C–I and I–M–C frameworks proposed by DynamicEarth (Li et al., 2026), as well as Seg2Change (Su et al., 2026), OpenDPR (Guo et al., 2026), AdaptOVCD (Dou et al., 2026), OmniOVCD (Zhang et al., 2026), and ReA-OVCD (Zhu et al., 2026).

4.3.1 Results on Building Change Detection

As shown in Table 1, BiRNet achieves competitive performance on LEVIR-CD, WHU-CD, and S2Looking, obtaining the best results on LEVIR-CD and S2Looking. Particularly, on S2Looking, which contains significant viewpoint variations and complex backgrounds, BiRNet achieves a substantial improvement over the previous best method, demonstrating its ability to effectively suppress false alarms caused by appearance variations and background interference. Moreover, BiRNet achieves the highest Precision on all three datasets, verifying its capability to filter out unreliable change responses. As shown in Fig. 4(a), existing methods tend to generate missed detections and false alarms in complex scenes, whereas BiRNet produces more complete change regions with fewer erroneous responses. In contrast, WHU-CD contains relatively simple and regular building categories, where semantic information provides limited additional constraints, resulting in relatively smaller performance gains.

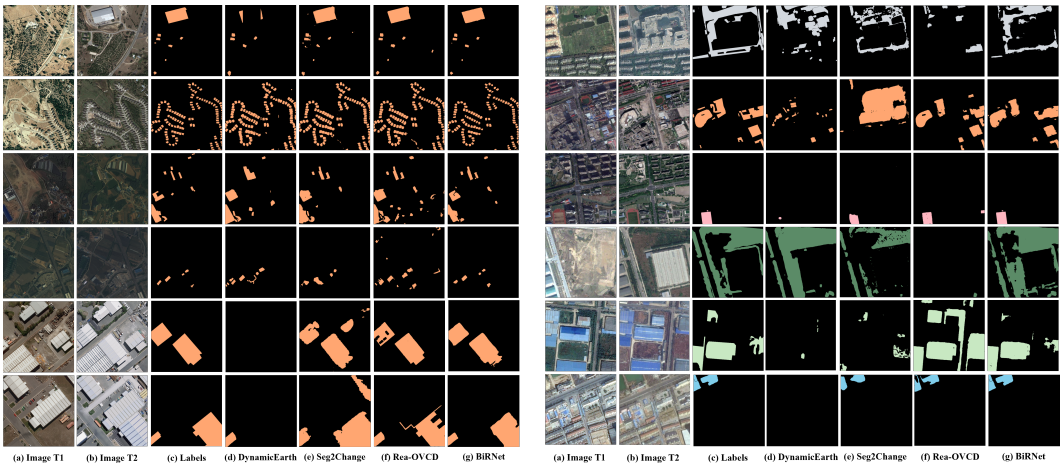

Figure 4: Qualitative results of BiRNet. (a) Building change detection results on LEVIR-CD, WHU-CD, and S2Looking. (b) Land-cover change detection results on SECOND. Color rendering: orange: "building", blue: "water", pink: "playground", light green: "low vegetation", dark green: "tree", and gray: "bareland".

Table 3: Plug-and-play validation of BCCST. The best and second-best results are marked in bold and underline, respectively.

Method	LEVIR-CD		WHU-CD		S2Looking		SECOND	
	IoU^C	$F1^C$	IoU^C	$F1^C$	IoU^C	$F1^C$	mIoU	mF1
Seg2Change	**64.89**	**78.71**	**75.46**	**86.02**	**21.05**	**34.77**	**29.01**	**42.78**
Seg2Change*	44.37	61.46	60.13	75.10	<u>20.26</u>	<u>33.69</u>	26.17	38.40
Seg2Change*+BCCST	<u>46.35</u>	<u>63.34</u>	<u>63.46</u>	<u>77.65</u>	19.84	33.11	<u>28.13</u>	<u>41.59</u>

4.3.2 Results on Land-Cover Change Detection

As shown in Table 2, BiRNet improves the mean IoU^C and $F1^C$ on the SECOND dataset by 6.8% and 6.93%, respectively, demonstrating that jointly leveraging change information and bi-temporal semantic information can effectively enhance semantic change recognition in open-vocabulary scenarios. From the category perspective, performance variations remain among different land-cover types. Specifically, the Playground category exhibits regular spatial structures and distinctive semantic features, enabling more stable change localization and category recognition. In contrast, natural categories such as Low vegetation have higher visual similarity with other categories and ambiguous semantic boundaries, making them more challenging to detect. As shown in Fig. 4(b), BiRNet generates more accurate semantic change predictions and reduces incorrect responses compared with existing methods.

4.4 Plug-and-Play Validation of BCCST

To verify the plug-and-play capability of BCCST, we construct Seg2Change* and Seg2Change*+BCCST. Seg2Change* replaces the original OVSS branch with CAFe-DINO used in BiRNet, while Seg2Change*+BCCST directly introduces BCCST without architecture modification or additional training. As shown in Table 3, Seg2Change* consistently underperforms Seg2Change on all datasets, indicating that the improvement of BiRNet mainly comes from the effective interaction between change information and semantic information rather than simply relying on a stronger semantic model. After incorporating BCCST, Seg2Change*+BCCST achieves further improvements on LEVIR-CD, WHU-CD, and SECOND, demonstrating the effectiveness of change-constrained semantic transition. The slight degradation on S2Looking is attributed to its complex scenes and large appearance variations, which weaken the complementarity between change and semantic probabilities.

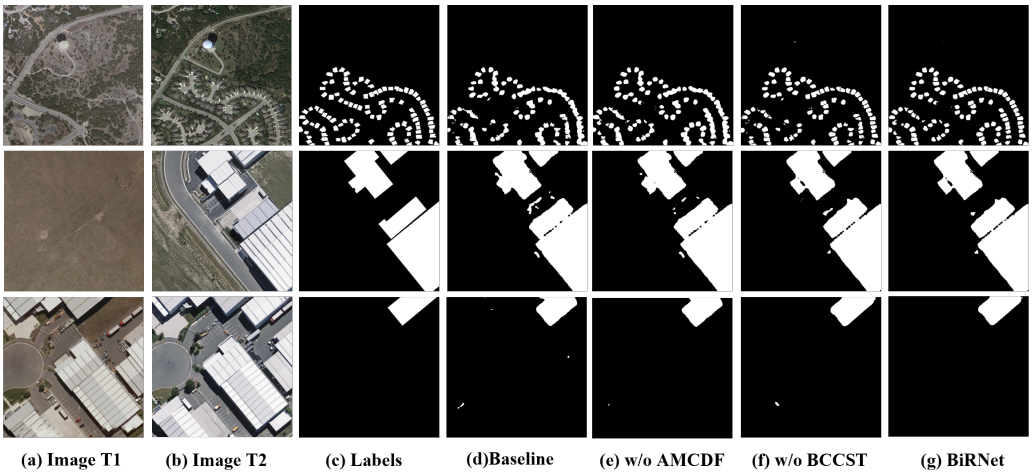

Figure 5: Qualitative comparison of BiRNet variants on building change detection datasets.

Table 4: Ablation study of AMCDF and BCCST. We report IoU^C and F1^C on three building change detection datasets, together with class-average results on SECOND.

Method	LEVIR-CD		WHU-CD		S2Looking		SECOND	
	IoU^C	F1^C	IoU^C	F1^C	IoU^C	F1^C	mIoU	mF1
Baseline	65.20	78.93	79.53	88.60	40.64	57.79	36.99	52.53
w/o BCCST	<u>73.69</u>	<u>84.85</u>	**80.68**	**89.31**	<u>45.00</u>	<u>62.07</u>	<u>38.10</u>	<u>53.57</u>
w/o AMCDF	66.69	80.02	78.89	88.20	41.67	58.82	37.06	52.57
BiRNet	**75.85**	**86.27**	<u>80.46</u>	<u>89.17</u>	**46.26**	**63.26**	**38.28**	**53.69**

4.5 Ablation Study

To evaluate the effectiveness of AMCDF and BCCST, we conduct ablation experiments by removing each module separately. The quantitative results are reported in Table 4, while the visualization results are shown in Figure 5. BiRNet achieves the best overall performance on most datasets.

Effect of AMCDF. Removing AMCDF degrades performance across all datasets. As shown in Figure 5, the variant produces more false detections and incomplete boundaries, demonstrating its effectiveness in enhancing change representation.

Effect of BCCST. Removing BCCST causes performance drops on LEVIR-CD, S2Looking, and SECOND. The limited improvement on WHU-CD is attributed to regular buildings and reliable change localization, where semantic information provides limited constraints. Visualization results show that BCCST suppresses false detections and improves changed-region completeness.

5 Conclusion

This paper proposes BiRNet, a unified bidirectional reasoning paradigm for open-vocabulary remote sensing change detection. First, we point out that existing OVCD methods lack effective interaction between change localization and semantic recognition, which limits their capabilities.To address this issue, BiRNet jointly models binary change probabilities and bi-temporal semantic information, enabling both change information and semantic information to participate in semantic change reasoning. Furthermore, an Adaptive Multi-scale Context-Difference Fusion Module (AMCDF) is designed to enhance change representation, and a Bidirectional Change-Constrained Semantic Transition Module (BCCST) is proposed to constrain semantic transitions using change probabilities, achieving unified prediction of change states and semantic categories. Finally, BiRNet achieves competitive performance on multiple OVCD benchmarks.

AI USE STATEMENT

Generative AI tools were used only to assist with language editing and polishing. They were not used for literature searches, research ideation or execution, drafting substantive manuscript paragraphs, synthetic data generation, or mathematical proof development; other AI uses are not applicable to this work. We reviewed the AI-assisted edits and take responsibility for the final content of this work, including its text, claims, and results.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, pp. 138846–138923, 2026.

Kaiyu Li, Xiangyong Cao, Yupeng Deng, Jiayi Song, Junmin Liu, Deyu Meng, and Zhi Wang. Semicd-vl: Visual-language model guidance makes better semi-supervised change detector. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision*, pp. 9630–9640, 2021.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020.

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10545–10556, 2025a.

Mingyu Dou, Shi Qiu, Ming Hu, Yifan Chen, Huping Ye, Xiaohan Liao, and Zhe Sun. Adaptovcd: Training-free open-vocabulary remote sensing change detection via adaptive information fusion. *arXiv preprint arXiv:2602.06529*, 2026.

Ryan Faulkenberry and Saurabh Prasad. Dino soars: Dinov3 for open-vocabulary semantic segmentation of remote sensing imagery. *arXiv preprint arXiv:2605.03175*, 2026.

Almog Friedlander, Ariel Shamir, and Ohad Fried. Goldilocs: General object-level detection and labeling of changes in scenes. In *International Conference on Learning Representations*, pp. 51033–51056, 2026.

Kaiyu Li, Shengqi Zhang, Yujie Wang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images. *arXiv preprint arXiv:2512.08730*, 2025b.

Junyu Gao, Da Zhang, Feiyu Wang, Lichen Ning, Zhiyuan Zhao, and Xuelong Li. Combining sam with limited data for change detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. doi: 10.1109/TGRS.2025.3545040.

Qi Guo, Jue Wang, Yinhe Liu, and Yanfei Zhong. Opendpr: Open-vocabulary change detection via vision-centric diffusion-guided prototype retrieval for remote sensing imagery. *arXiv preprint arXiv:2603.27645*, 2026.

Shiqi Huang, Shuting He, Huaiyuan Qin, and Bihan Wen. Score: Scene context matters in open-vocabulary remote sensing instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12559–12569, 2025.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24905–24916, 2025.

Dong Un Kang, Hayeon Kim, and Se Young Chun. Class distribution-induced attention map for open-vocabulary semantic segmentations. In *International Conference on Learning Representations*, pp. 12189–12212, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 3992–4003, 2023. doi: 10.1109/ICCV51070.2023.00371.

Zuzheng Kuang, Honghao Chang, Boqiang Liang, Haoqian Wang, Lijun He, Fan Li, and Haixia Bi. Memovcd: Training-free open-vocabulary change detection via cross-temporal memory reasoning and global-local adaptive rectification. *arXiv preprint arXiv:2604.26774*, 2026.

Kaiyu Li, Xiangyong Cao, Yupeng Deng, Chao Pang, Zepeng Xin, Hui Qiao, Tieliang Gong, Deyu Meng, and Zhi Wang. Dynamicearth: How far are we from open-vocabulary change detection? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6279–6287, 2026.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 70–88, 2024.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *International Conference on Learning Representations*, pp. 17904–17925, 2024.

Yuxiang Lu, Shengcao Cao, and Yu-Xiong Wang. Swiss army knife: Synergizing biases in knowledge from vision foundation models for multi-task learning. In *International Conference on Learning Representations*, pp. 80538–80574, 2025.

Maximilian Nentwich, Marcin Leszczynski, Brian E. Russ, Lukas Hirsch, Noah Markowitz, Kaustubh Sapru, Charles E. Schroeder, Ashesh D. Mehta, Stephan Bickel, and Lucas C. Parra. Semantic novelty modulates neural responses to visual change across the human brain. *Nature Communications*, 14(1):2910, 2023.

Xiaogang Ning, Hanchao Zhang, Ruiqian Zhang, and Xiao Huang. Multi-stage progressive change detection on high resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 207:231–244, 2024. doi: 10.1016/j.isprsjprs.2023.11.023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Sunghyun Park, Jungsoo Lee, Shubhankar Borse, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. Understanding personal concept in open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19957–19966, 2025.

Daifeng Peng, Yuanke Peng, and Haiyan Guan. Zero-ovcd: Bridging training-free foundation models and pseudo-label learning for open-vocabulary change detection. *arXiv preprint arXiv:2608.11663*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Girish Agarwal, Sandhini Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021.

Cheng Shi and Sibei Yang. The devil is in the object boundary: Towards annotation-free instance segmentation using foundation models. In *International Conference on Learning Representations*, pp. 12790–12815, 2024.

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23487–23497, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Sara Spotorno, Benjamin W. Tatler, and Sylvane Faure. Semantic consistency versus perceptual salience in visual scenes: Findings from change detection. *Acta Psychologica*, 142(2):168–176, 2013.

You Su, Yonghong Song, Jingqi Chen, and Zehan Wen. Seg2change: Adapting open-vocabulary semantic segmentation model for remote sensing change detection. *arXiv preprint arXiv:2604.11231*, 2026.

Weidong Tang, Hanbin Sun, Zihan Li, Yikai Wang, and Feifan Zhang. Coregovcd: Consistency-regularized open-vocabulary change detection. *arXiv preprint arXiv:2604.02160*, 2026.

Zhengyang Wang, Tingliang Feng, Fan Lyu, Fanhua Shang, Wei Feng, and Liang Wan. Dual semantic guidance for open vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20212–20222, 2025.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations*, pp. 52484–52503, 2024.

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:3113912, 2022.

Xinyu Zhang, Haihua Chen, Yanzhi Zhao, Ming He, and Xiaoqing Han. Change detection of buildings in remote sensing images using a spatially and contextually aware siamese network. *Expert Systems with Applications*, 276:127110, 2025.

Xu Zhang, Danyang Li, Yingjie Xia, Xiaohang Dong, Hualong Yu, Jianye Wang, and Qicheng Li. Omniovcd: Streamlining open-vocabulary change detection with sam 3. *arXiv preprint arXiv:2601.13895*, 2026.

Maofan Zhao, Xinli Hu, Linlin Zhang, Qingyan Meng, Yuxing Chen, and Lorenzo Bruzzone. Beyond pixel-level annotation: Exploring self-supervised learning for change detection with image-level supervision. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3379431.

Hongming Zhu, Huaji Chen, Bowen Du, Sicong Liu, and Qin Liu. Rea-ovcd: Reliability-aware open-vocabulary change detection via semantic and spatial refinement. *arXiv preprint arXiv:2606.20032*, 2026.

Ziqiang Zhu and Bowei Yang. Univcd: A new method for unsupervised change detection in the open-vocabulary era. *arXiv preprint arXiv:2512.13089*, 2025.

Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary high-resolution remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.