Free-CD: Probabilistically Decoupled Training-Free Open-Vocabulary Change Detection with Resolution-Invariant Feature Inversion

Yongshuo Zhu[1], Lu Li[1,2,*], Keyan Chen[1,3], Zhenwei Shi[1,3], and Fugen Zhou[1]

[1] Department of Aerospace Intelligent Science and Technology, School of Astronautics, Beihang University, Beijing 100191, China
[2] State Key Laboratory of High-Efficiency Reusable Aerospace Transportation Technology, Beijing, 102206, China
[3] Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technologies, Ministry of Education, China
lilu@buaa.edu.cn

Abstract. Open-Vocabulary Change Detection (OVCD) faces a fundamental granularity gap: foundation models prioritize high-level semantic abstraction while change detection requires pixel-level spatial fidelity. Traditional OVCD methods rely on instance extraction models, introducing spatial semantic ambiguities and leading to over-segmentation or under-segmentation in remote sensing scenarios. To bridge this gap, we propose **Free-CD**, a training-free OVCD framework that reformulates the task by predicting a change probability distribution rather than enforcing binary change masks via instance boundaries. We introduce RIFI-Up, an upsampler that enforces a resolution-invariant feature inversion constraint, enabling high-fidelity detail recovery without semantic artifacts. Additionally, we propose a Bayesian Probability Correction mechanism that separates change priors from semantic categorization, introducing a correction term to rectify change area discrimination and enforce bi-temporal semantic consistency. Extensive experiments on four remote sensing datasets demonstrate that Free-CD consistently outperforms current methods, establishing new standards in accuracy and generalization for open-vocabulary change detection. Code is available at https://github.com/20374230/FreeCD.

Keywords: Remote sensing · Change detection · Open vocabulary

1 Introduction

Semantic Change Detection (SCD) stands at the critical intersection of Earth observation and computer vision, transforming bi-temporal data into action-

* corresponding author.

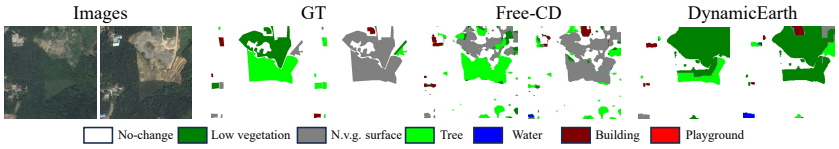

Fig. 1: A qualitative comparison between DynamicEarth [29] and the proposed Free-CD framework. The results illustrate that DynamicEarth exhibits clear limitations in both region delineation and category discrimination accuracy.

able insights for urban planning, disaster response, and environmental conservation [9,12,13,54,62]. While supervised learning has achieved pixel-level precision, it remains restricted by fixed vocabularies. The dynamic nature of the physical world demands Open-Vocabulary Change Detection (OVCD), a paradigm leveraging Vision-Language Models (VLMs) to identify arbitrary changes beyond predefined categories. However, the core challenge of OVCD lies in a fundamental granularity gap: foundation models (e.g., CLIP, DINO) prioritize high-level coarse semantic abstraction, whereas change detection necessitates pixel-level spatial fidelity.

To bridge this granularity gap, contemporary OVCD architectures like DynamicEarth [29] leverage instance extraction models like the SAM [27] or APE [46] to provide fine-grained spatial boundaries. However, this paradigm tightly couples detection with recognition and imposes rigid semantic boundaries, overlooking a critical fact: the instance extraction process itself inherently introduces spatial semantic ambiguities. This manifests as over-segmentation (splitting a single semantic object into multiple masks) or under-segmentation (merging distinct objects into one). In remote sensing scenarios, where region-based land cover types (e.g., farmland, grasslands) significantly outnumber discrete object instances (e.g., buildings), such low-quality instance extraction severely corrupts semantic discrimination. As illustrated in Fig. 1, this architecture proves particularly fragile when handling land cover classes with strong regional continuity (e.g., bare soil) or fuzzy semantic boundaries (e.g., low vegetation and trees). Crucially, forcibly generating deterministic masks and assigning them singular semantics introduces spurious confidence. This underscores the need for a mechanism that can explicitly model boundary fuzziness and semantic uncertainty.

Motivated by these observations, we reformulate the OVCD objective. Instead of enforcing binary change masks via instance boundaries, we recast potential change region proposals as predictions of a change probability distribution. This probabilistic formulation naturally accommodates ambiguity while maintaining pixel-level spatial fidelity through a high-fidelity feature upsampling scheme. Acknowledging the scarcity of annotated data in open-vocabulary settings, we explore a training-free pathway that leverages the intrinsic priors of foundation models to mine latent change relations in remote sensing scenes. Specifically, we propose Free-CD, a training-free OVCD framework that decomposes the task into two parallel sub-tasks: potential change region proposal

and bi-temporal open-vocabulary segmentation. To ensure consistency between change detection and semantic discrimination, we introduce a Bayesian Probability Correction mechanism. This mechanism jointly calibrates the change probability map using bi-temporal semantic cues and enforces bi-temporal consistency via robust mask operations. Furthermore, to address semantic artifacts caused by traditional upsampling, we design RIFI-Up. This novel module incorporates a resolution-invariant feature inversion constraint, ensuring high-fidelity spatial detail recovery while preserving semantic consistency throughout the upsampling process. The contributions of this paper are summarized as follows:

i) Free-CD: A training-free OVCD framework that mathematically reformulates spatial upscaling and semantic inference, enabling high-performance change detection across multiple categories without task-specific training.

ii) RIFI-Up: A feature upsampler enforcing a resolution-invariant feature inversion constraint. It enables pixel-level fidelity through direct reconstruction without semantic drift by leveraging resolution-invariance from Fourier Neural Operators.

iii) Bayesian Probability Correction: A mechanism that separates change priors from semantic categorization, introduces a correction term to rectify change area discrimination, and enforces bi-temporal semantic consistency.

iv) SOTA Performance: Extensive experiments on four remote sensing benchmarks demonstrate Free-CD's superiority over state-of-the-art methods, establishing new standards in accuracy and generalization for open-vocabulary change detection.

2 Related Works

2.1 Remote Sensing Image Change Detection

Remote sensing image change detection is primarily categorized into two tasks: Binary Change Detection (RSIBCD) and Semantic Change Detection (RSISCD) [41]. RSIBCD identifies significant surface changes between bi-temporal images, outputting a binary change map. Most deep learning-based methods employ convolutional or transformer architectures for supervised learning, focusing on enhancing feature representation and discrimination. Research directions include minimizing information loss [20, 35, 42], designing feature enhancement modules [3, 15, 24, 61], and leveraging edge information [2, 31, 52]. Recent efforts also explore adapting foundational models like SAM for change detection [10].

RSISCD extends binary detection by further identifying the specific types of changes (e.g., from forest to urban). Mainstream approaches include: Multi-task Learning Frameworks, which jointly solve binary change detection and semantic segmentation [8, 16, 17, 23, 38]. These works focus on spatiotemporal consistency, multi-scale fusion, and task-decoupling strategies. Post-Classification Comparison (PCC) compares independent semantic segmentation results from each timestamp [16, 50]. While prone to error accumulation, this paradigm has inspired recent OVCD methods like DynamicEarth [29]. In summary, despite

significant progress, the field still faces challenges, including limited quality and scale of the labeled dataset, the complexity of detecting diverse land cover changes in high-resolution imagery, and the growing demand for algorithmic efficiency in practical applications. OVCD thus presents a novel and transformative solution. Its ability to transcend predefined categories and operate with user-defined semantic queries directly addresses the core limitations of conventional methods, offering a path toward more generalizable and annotation-efficient change analysis.

2.2 Open-Vocabulary Semantic Segmentation

Open-Vocabulary Semantic Segmentation (OVSS) aims to segment images based on arbitrary textual concepts, breaking away from the constraints of predefined categories. Early OVSS methods predominantly adopted two-stage frameworks [11, 22, 25, 32, 34, 43, 55, 56], which first utilize instance proposal networks (e.g., Mask2Former [14] or SAM [27]) to extract category-agnostic mask candidates, followed by open-vocabulary classification using vision-language models such as CLIP [44]. While effective in natural images, the two-stage OVSS paradigm faces two domain-specific challenges when applied to remote sensing that hinder its effectiveness. First, under high resolution, instance proposers like SAM generate excessive fine-grained segments, which leads to over-segmentation and introduces numerous trivial instances that severely interfere with subsequent semantic discrimination. Second, ambiguous object boundaries arise because many land cover types (e.g., low vegetation and trees) exhibit gradual transitions and fuzzy semantic boundaries, which confuses instance extraction and leads to fragmented or inaccurate proposals. To overcome these limitations, recent remote sensing OVSS methods have increasingly adopted one-stage frameworks [5, 28, 53, 58]. These approaches directly predict semantic masks from image features using a unified architecture, eliminating the reliance on standalone instance proposal networks. In handling diverse remote sensing scenarios under the OVSS setting, one-stage methods circumvent the problematic instance extraction step, thereby effectively avoiding over-segmentation and boundary ambiguity, and demonstrating stronger robustness and better accuracy.

Recent advances in remote sensing-oriented OVSS methods [5,28,30,58] have demonstrated strong generalization capability and segmentation accuracy across diverse remote sensing scenarios and multiple land cover categories. Inspired by these works, FreeCD abandons instance-level change proposals, reinterprets change detection as the prediction of change probability distributions, and yields promising performance.

3 Task Definition and Formulation

Let F denote the event that a change occurs at a given location between bi-temporal remote sensing images, and let $\mathcal{C}$ represent the predefined open-vocabulary category set (including both foreground categories and the background class).

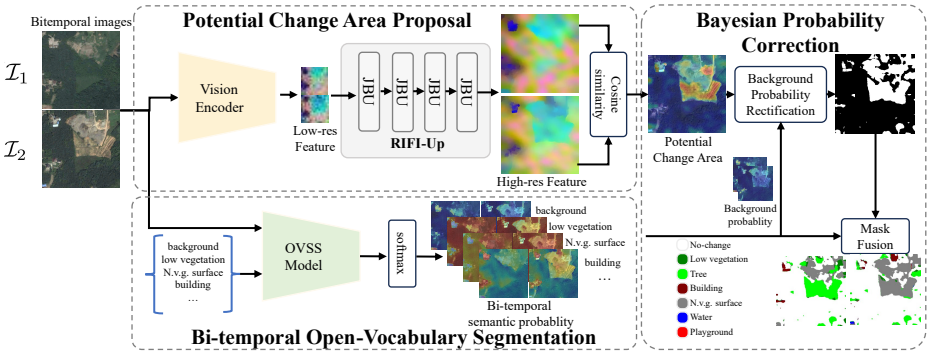

Fig. 2: Illustration of Free-CD Framework. The processing pipeline consists of three steps: Potential Change Area Proposal, Bi-temporal Open-Vocabulary Segmentation, and Bayesian Probability Correction.

One of the OVCD objectives is to predict the joint conditional probability $P(F, c_i \mid \mathcal{C})$ of a location undergoing a change belonging to category $c_i \in \mathcal{C}$ in one temporal phase. Since the change event F depends solely on regional differences between bi-temporal images and is independent of specific categories, this joint probability can be decomposed as:

$$P(F, c_i \mid \mathcal{C}) = P(F) \cdot P(c_i \mid \mathcal{C}), \tag{1}$$

where $P(F)$ represents the probability of a potential change occurring in the region (corresponding to the Potential Change Area Proposal task), and $P(c_i \mid \mathcal{C})$ denotes the conditional probability of the region belonging to category c_i (corresponding to the Bi-temporal Open-Vocabulary Segmentation task).

4 Free-CD

In this paper, we present Free-CD, a unified OVCD framework as illustrated in Fig. 2. This framework separates OVCD into two parallel tasks: potential change area proposal and bi-temporal open-vocabulary segmentation. The potential change area proposal process estimates the probability $P(F)$ of a region undergoing a potential change. The bi-temporal open-vocabulary segmentation branch performs object-boundary segmentation and category discrimination on the bi-temporal images, yielding the conditional probability $P(c_i \mid \mathcal{C})$ for each category $c_i \in \mathcal{C}$ for each time phase. In the Bayesian Probability Correction stage, the results from the two branches are integrated through background-probability correction and mask fusion, producing the final change-mask image.

4.1 Potential Change Area Proposal

We adopt a straightforward approach to obtain the potential change region proposal: the features extracted by the feature encoder from the bi-temporal images

are directly up-sampled, and then the similarity of the high-resolution features is compared pixel-wise to obtain the potential change area probability $P(F)$. This process can be formally described as follows:

$$\begin{aligned} F_i &= \Phi_{\text{enc}}(\mathcal{I}_i), \quad i \in \{1,2\}, \\ F_i^{\text{up}} &= \Phi_{\text{up}}(F_i), \quad i \in \{1,2\}, \\ P(F) &= \Phi_{\text{sigmoid}}\big(-\Phi_{\text{sim}}(F_1^{\text{up}}, F_2^{\text{up}})\big). \end{aligned} \tag{2}$$

Given bi-temporal remote sensing images $\mathcal{I}_i, i \in \{1,2\}$, the pre-trained feature encoder Φ_{enc} first extracts low-resolution features F_i. Then, the feature upsampler RIFI-Up $\Phi_{\text{up}}(\cdot)$ is applied to obtain high-resolution bi-temporal features F_i^{up}. RIFI-Up leverages the stacked parameterized Joint Bilateral Upsampler (JBU) [21] to operate feature upsampling. Details of RIFI-Up can be found in Sec. 5. Since the JBU operates with reference to the original image resolution, the potential change region proposal process is robust to resolution variations. Following [60], cosine similarity is used to measure the degree of change between bi-temporal images, and a sigmoid function $\Phi_{\text{sigmoid}}(\cdot)$ normalizes the cosine values into a probability representation.

4.2 Bi-temporal Open-Vocabulary Segmentation

The Bi-temporal Open-Vocabulary Segmentation process performs semantic segmentation independently on each temporal image to delineate object boundaries and identify semantic categories in the bi-temporal imagery. Given a category set $\mathcal{C}$, the OVSS model outputs, for each pixel in a single-temporal image, the per-class conditional probability $P(c_j \mid \mathcal{C})$ for every category $c_j \in \mathcal{C}$. This can be abstractly formulated as:

$$P_i(c_j \mid \mathcal{C}) = \Phi_{\text{OVSS}}(\mathcal{I}_i, c_j), \quad i \in \{1,2\},\ c_j \in \mathcal{C}, \tag{3}$$

where $P_i(c_j \mid \mathcal{C})$ denotes the per-pixel conditional probability for category c_j in the image of temporal phase i, and $\Phi_{\text{OVSS}}(\cdot)$ represents any arbitrary OVSS model.

4.3 Bayesian Probability Correction

The Bayesian Probability Correction module operates through a two-stage pipeline: **background probability rectification** and **mask fusion**. First, we correct the background probabilities to accurately derive the conditional likelihood that a change proposal belongs to foreground categories, addressing definitional discrepancies between single- and multi-temporal settings. Subsequently, mask fusion integrates these calibrated probabilities to generate the final semantic change masks.

Background probability rectification: To ensure probabilistic completeness, the category set $\mathcal{C}$ must form a complete event group, satisfying $\sum_{c_i \in \mathcal{C}} P(c_i \mid$

$\mathcal{C}) = 1$. Typically, $\mathcal{C}$ includes a "background" class B indicating that a pixel does not belong to any foreground category. However, in single-temporal semantic segmentation, the background probability $P(B \mid \mathcal{C})$ reflects the event that a "pixel does not belong to any foreground category," whereas in multi-temporal change detection, the event F' "change belongs to the background" constitutes the union of "no change occurred" and "the change does not belong to any foreground category". This definitional discrepancy introduces bias in bi-temporal background discrimination, leading to inconsistent OVCD results. To address this, we introduce a background correction mechanism. Specifically, let $V = \mathcal{C} \setminus \{\text{"background"}\}$ denote the set of foreground categories, and let event B represent "a region is predicted as background in a single-temporal image." Thus, the corrected probability of event F' ("change occurs and belongs to the foreground class") for the predefined category set $\mathcal{C}$ in multi-temporal settings is defined as:

$$P(F' \mid \mathcal{C}) = P(F) \cdot (1 - \prod_{i=1}^{2} P_i(B \mid \mathcal{C})), i \in \{1, 2\}, \tag{4}$$

where $P_i(B \mid \mathcal{C})$ represents the conditional probability of a region being predicted as background in temporal phase i. With this correction, the OVSS task only needs to classify foreground regions, and the OVCD objective is reformulated as:

$$P(F', c_i \mid \mathcal{C}) = P(F' \mid \mathcal{C}) \cdot P(c_i \mid V), \quad c_i \in V. \tag{5}$$

The Free-CD framework adopts this refined task definition, effectively resolving background-induced biases and ensuring consistent bi-temporal change region identification.

Mask Fusion: Following Equation 5, we generate the final change mask through a straightforward two-step procedure. First, a thresholding operation converts the probability $P(F' \mid \mathcal{C})$ into a binary state, indicating regions where a change has occurred. Subsequently, this binary mask is applied to the semantic probability maps $P_i(c_j \mid \mathcal{V})$ to assign semantic labels to the changed areas. This process is formalized as:

$$\begin{aligned} M_{\text{BCD}} &= \Phi_{\text{threshold}}\big(P(F' \mid \mathcal{C})\big) \\ M^i_{\text{SCD}} &= M_{\text{BCD}} \odot \Phi_{\text{argmax}}\big(P_i(c_j \mid \mathcal{V})\big), i \in \{1, 2\},\ c_j \in \mathcal{V}, \end{aligned} \tag{6}$$

where M_{BCD} denotes the binary change mask obtained via the thresholding operator $\Phi_{\text{threshold}}(\cdot)$. The semantic change mask for temporal phase i, denoted as M^i_{SCD}, is generated by first applying the argmax operator $\Phi_{\text{argmax}}(\cdot)$ to obtain the dominant foreground category at each pixel, followed by an element-wise multiplication ($\odot$) with M_{BCD}. This operation effectively sets regions classified as background to zero, resulting in a mask that highlights only the changed areas with their corresponding semantic labels.

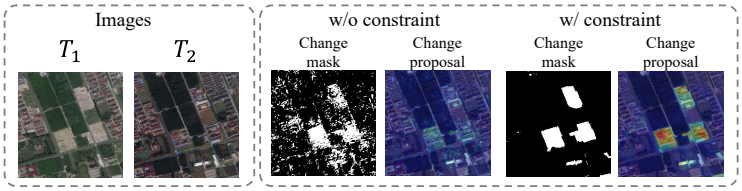

Fig. 3: Effect of the Resolution-Invariant Constraint. Shown are the bi-temporal images together with the corresponding change proposals and change masks generated with (w/) and without (w/o) the resolution-invariant constraint. Without the constraint, the proposals exhibit noticeable artifacts.

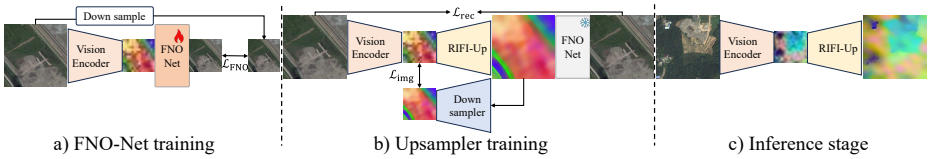

Fig. 4: Training and inference procedure of RIFI-Up. The training stage consists of two stages: a) FNO network training for feature-image alignment, and b) upsampler training to obtain RIFI-Up. Only RIFI-Up is used during the c) inference stage.

5 RIFI-Up

To mitigate semantic artifacts from traditional upsampling, we propose RIFI-Up, a feature upsampler that integrates a resolution-invariant feature inversion constraint. Leveraging the inherent resolution invariance of Fourier Neural Operators (FNO), RIFI-Up enforces a transformation that reconstructs the original image directly from upsampled features, ensuring high-fidelity detail recovery without artifacts.

5.1 Construction of the Encoder Inverse Mapping

Let the image space be $\mathcal{I} \subset L^2(\Omega; \mathbb{R}^3)$ and $\Phi_{\text{enc}} : \mathcal{I} \to \mathcal{F}$ an encoder network mapping an image I to a feature map $F = \Phi_{\text{enc}}(I)$. To design an upsampler that enhances features while preserving semantics, we seek an inverse mapping $\Phi_{\text{enc}}^{-1} : \mathcal{F} \to \mathcal{I}$ such that $I \approx \Phi_{\text{enc}}^{-1}(F)$.

Although ViT encoders are theoretically injective [37], their complex nonlinear form precludes analytic inversion. We therefore approximate the inverse by a learnable pseudo-inverse G_θ minimizing the reconstruction error:

$$\min_{G_\theta} \mathbb{E}_{I \sim \mathcal{I}} \left\| G_\theta\big(\Phi_{\text{enc}}(I)\big) - I \right\|^2 . \tag{7}$$

The existence of such an approximator is guaranteed by the universal approximation theorem for neural operators [1].

To apply the inverse mapping for feature upsampling, we introduce an upsampling operator $\Phi_{\text{up}} : \mathcal{F} \to \mathcal{F}_{\text{high}}$. The desired composite mapping is:

$$I \approx \Phi_{\text{enc}}^{-1}\big(\Phi_{\text{up}}(F)\big). \tag{8}$$

For the upsampler to produce high-fidelity features that generalize across scales, its inverse mapping must satisfy resolution invariance. That is, for low-resolution features $F = \Phi_{\text{enc}}(I)$ corresponding to the down-sampled image $\Phi_{\text{down}}(I)$, the same inverse mapping should accurately reconstruct that low-resolution image:

$$\Phi_{\text{down}}(I) \approx \Phi_{\text{enc}}^{-1}(F). \tag{9}$$

Without this constraint, the learned inverse mapping becomes merely a common fitting function from specific high-resolution features to the image, which is more sensitive to details in the input features. When processing multi-temporal features, such a mapping tends to distort normal feature differences between the two temporal phases, thereby introducing pronounced artifacts in the generated change proposals, as shown in Fig. 3. Resolution invariance forces the inverse mapping to remain stable and continuous in function space, decoupling it from the absolute scale of the input features and filtering out spurious noise introduced by resolution differences. This ensures semantic consistency and effectively avoids artifacts.

5.2 Implementing the Inverse Mapping with Fourier Neural Operators

FNOs learn resolution-invariant inverse mappings. An FNO layer operates in the Fourier domain:

$$v_{t+1}(x) = \sigma\left(W v_t(x) + \mathcal{F}^{-1}\big(R_\phi \cdot \mathcal{F}(v_t)\big)(x)\right), \tag{10}$$

where $\mathcal{F}$ is the Fourier transform, R_ϕ a learnable spectral weight matrix, W a linear transformation, and σ a non-linear activation. The spectral parameterization makes the operator grid-independent, enabling evaluation at any resolution after training on one.

We instantiate the inverse mapping as FNO $\mathcal{G}_\theta : \mathcal{F} \to \mathcal{I}$. It is trained on a low-resolution dataset $\mathcal{D}_l = \{(F_l, I_l)\}$ where $F_l = \Phi_{\text{enc}}(I_l)$ and $I_l = \Phi_{\text{down}}(I)$ for high-resolution images I, minimizing:

$$\min_\theta \ \mathbb{E}_{(F_l, I_l)\sim\mathcal{D}_l} \left\|\mathcal{G}_\theta(F_l) - I_l\right\|^2. \tag{11}$$

After training, $\mathcal{G}_{\theta^*}$ satisfies resolution invariance: for any spatial domain Ω, $\mathcal{G}_{\theta^*}(F|_\Omega) = \mathcal{G}_{\theta^*}(F)|_\Omega$. Given high-resolution features F_h, it directly produces $\hat{I}_h = \mathcal{G}_{\theta^*}(F_h) \in \mathcal{I}_h$. This yields a continuous, resolution-invariant approximation $\Phi_{\text{enc}}^{-1} \approx \mathcal{G}_{\theta^*}$, used to build the upsampler RIFI-Up.

5.3 Training Procedure

The training consists of two stages: FNO network training and upsampler training, as shown in Fig. 4

1) FNO Network Training Stage: This stage uses low-resolution images and their corresponding features to train the FNO network for feature-to-image alignment. The loss function is defined as:

$$\mathcal{L}_{\text{FNO}} = \left\| F - \Phi_{\text{down}}(I) \right\|_2^2, \tag{12}$$

where F represents the low-resolution features, $\Phi_{\text{down}}(I)$ is the downsampled original image, and $\| \cdot \|_2$ denotes the squared l2 norm.

2) Upsampler Training Stage: Following FeatUp [21], the upsampler training is governed by a multi-view reconstruction loss $\mathcal{L}_{\text{rec}}$ and an image restoration constraint $\mathcal{L}_{\text{img}}$:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \left\| \Phi_{\text{enc}}(t(I)) - \Phi'_{\text{down}}\big(t(\Phi_{\text{up}}(F))\big) \right\|_2^2 \\ \mathcal{L}_{\text{img}} &= \left\| I - \Phi_{\text{enc}}^{-1}\big(\Phi_{\text{up}}(F)\big) \right\|_2^2 \\ \mathcal{L} &= \mathcal{L}_{\text{rec}} + \gamma \cdot \mathcal{L}_{\text{img}}, \end{aligned} \tag{13}$$

where $\Phi'_{\text{down}}(\cdot)$ is a learnable downsampling network, $t(\cdot)$ denotes random perturbations (such as padding, zooming, cropping, horizontal flipping, and their compositions), and γ is a weighting coefficient. The parameters of the FNO network are frozen during this training phase.

Through this two-stage training, we obtain RIFI-Up, a feature upsampler with high fidelity and resolution robustness, which effectively supports feature comparison and region proposal in subsequent change detection tasks.

6 Experimental Results and Analyses

6.1 Experimental Setup

Implementation. The proposed model is implemented in PyTorch and runs on NVIDIA A100 GPUs. For the Potential Change Area Proposal module, we employ DINO-series models [6,39,47] as the feature backbone. To ensure a fair comparison, we adopt the ViT architecture as the backbone variant. For Bi-temporal Open-Vocabulary Segmentation, we adopt APE [46] and SegEarth-OV [30] for fair comparison. The RIFI-Up module is trained on the MillionAID dataset [36]. Both stages of the training process are carried out for 1 epoch with a batch size of 4, the weighting coefficient γ is 0.1. In the post-fusion stage, the Otsu thresholding method [40] is employed to separate changed areas from unchanged regions within the probability map.

Table 1: Comparison with the DynamicEarth framework on SECOND, HiUCD, HRSCD, and JL1-SCD. Best in Free-CD and best in DynamicEarth performances are highlighted.

Framework	Structure	SECOND		HiUCD		HRSCD		JL1-SCD	
		Fscd	mIoU	Fscd	mIoU	Fscd	mIoU	Fscd	mIoU
Dynamic-Earth	SAM-DINO-SegEarth-OV	16.02	50.37	20.86	48.79	2.81	43.82	7.86	38.20
	SAM-DINOv2-SegEarth-OV	18.91	56.22	22.67	52.35	3.12	48.02	7.87	38.58
	SAM-DINOv3-SegEarth-OV	20.50	58.77	23.24	53.21	3.18	48.38	8.07	40.31
	APE-/-DINO	4.97	50.88	14.00	48.73	0.58	32.23	5.58	43.83
	APE-/-DINOv2	4.99	51.11	18.08	52.96	2.89	34.73	10.23	44.45
	APE-/-DINOv3	4.74	50.70	17.96	53.45	3.65	35.70	10.26	44.72
Free-CD	DINOv1-SegEarth-OV	18.95	51.93	25.38	48.93	7.46	54.57	12.99	45.73
	DINOv2-SegEarth-OV	21.43	57.75	26.42	53.81	8.18	55.18	13.99	46.16
	DINOv3-SegEarth-OV	23.19	59.84	30.00	54.99	8.37	57.73	16.15	52.37
	DINOv1-APE	5.01	51.33	18.53	49.19	2.64	34.70	10.10	44.27
	DINOv2-APE	4.97	51.10	18.08	52.94	2.88	34.73	10.22	44.43
	DINOv3-APE	4.88	50.79	18.12	53.52	3.75	35.79	10.34	44.74

Dataset and Evaluation Metrics. The experiments are conducted on four commonly used SCD datasets: SECOND [57], HiUCD [51], HRSCD [16], and JL1-SCD, and two building change detection datasets: LEVIR-CD [10] and WHU-CD [26]. For the SCD task, following common SCD methods like [8], we adopt the Fscd score and mean Intersection over Union (mIoU) as evaluation metrics. For the building change detection task, following [29], F1-score and IoU are adopted.

6.2 Comparison with the State-of-the-Art

Results on Semantic Change Detection. We conducted comparative experiments between our proposed Free-CD and several open-vocabulary change detection (OVCD) frameworks from DynamicEarth [29]. Free-CD outperforms the M-C-I and I-M-C architectures of DynamicEarth across all datasets (Tab. 1). On the fine-grained HiUCD dataset, it achieves a higher Fscd score, demonstrating improved class discrimination. For the high-resolution HRSCD dataset, gains in both Fscd and mIoU reflect its robustness to scale- and region-level changes. Stable improvements are also observed on the diverse SECOND and JL1-SCD datasets, confirming the general effectiveness of the framework. Additionally, it is observed that under the APE architecture, the performance gap between Free-CD and DynamicEarth is relatively small, suggesting that the current performance bottleneck of the model is mainly attributed to the inferior identifier.

Qualitative experiments further validate these findings. As shown in Fig. 5, the following observations can be made: 1) Free-CD achieves more accurate change region segmentation for complex-shaped land parcels; 2) In high-resolution scenarios, Free-CD is less susceptible to interference from small, fragmented objects and demonstrates higher recognition accuracy for large, contiguous land

Table 2: The best and second-best results in each column are highlighted. "—" denotes data missing. Comparator in DynamicEarth and Free-CD is set as DINO-v2

Method	LEVIR-CD		WHU-CD	
	IoU	F1	IoU	F1
CVA [4]	—	12.2	—	—
PCA-KM [7]	4.8	9.1	5.4	10.2
CNN-CD [19]	7.0	13.1	4.9	9.4
DSFA [18]	4.3	8.2	4.1	7.8
DCVA [45]	7.6	14.1	10.9	19.6
GMCD [49]	6.1	11.6	10.9	19.7
DINOv2+CVA [60]	—	17.3	—	—
AnyChange-H [60]	—	23.0	—	—
SCM [48]	18.8	31.7	18.6	31.3
DynamicEarth (M-C-I) [29]	36.6	53.6	40.6	57.7
DynamicEarth (I-M-C) [29]	50.0	66.7	61.1	75.8
FreeCD (SegEarth-OV)	41.9	59.1	48.6	64.6
FreeCD (APE)	50.5	67.1	58.9	74.7

cover patches; 3) The boundaries between semantically similar categories are distinguished more clearly.

Results on Building Change Detection. Tab. 2 summarizes the results on building change detection datasets. FreeCD outperforms or matches existing methods on both benchmarks. Comparing the two Identifier settings, SegEarth-OV yields substantial improvements, while APE shows no gain. Buildings have distinct instance properties and require strong instance-level discrimination; APE possesses sufficient instance-awareness, approaching the task upper bound, so further framework modifications offer limited benefit. In contrast, SegEarth-OV is weaker on instance discrimination, and the framework effectively compensates for its deficiency, leading to notable gains. This indicates that the performance bottleneck of the APE setting lies primarily in the Identifier design rather than in the change detection framework itself, which aligns with observations from the SCD results.

6.3 Discussion on RIFI-Up Setup

We conduct ablation studies on different upsampling strategies in the potential change region proposal stage, with quantitative results reported in Tab. 3. The JBU variant enhanced with both feature inversion (FI) and resolution-invariant (RI) constraints—denoted as RIFI-Up—consistently achieves the best performance across all evaluated datasets. This confirms that RIFI-Up effectively improves change detection accuracy through high-fidelity feature recovery and adaptive resolution handling. Notably, even when using entirely training-free upsamplers such as bilinear interpolation or plain JBU, the framework still maintains a reasonable level of open-vocabulary change detection capability, underscoring the flexibility and generalizability of the proposed Free-CD architecture.

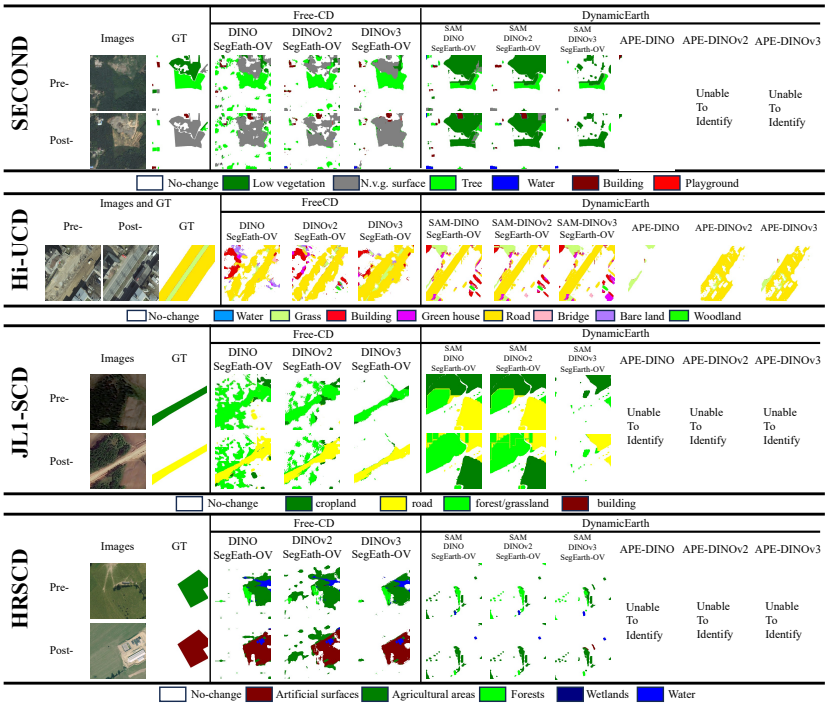

Fig. 5: Qualitative comparison results with other methods from DynamicEarth [29] on four datasets. Free-CD achieves more accurate change region segmentation for complex-shaped land parcels, is less susceptible to interference from small objects in high-resolution scenarios, demonstrates higher recognition accuracy for large, contiguous land cover patches, and distinguishes boundaries between semantically similar categories more clearly.

6.4 Framework Adaptability

Free-CD enables plug-and-play integration of any feature backbone with any OVSS model, directly converting an OVSS system into a full OVCD pipeline without architectural modifications or fine-tuning when backbones are shared. To validate this, we evaluate three CLIP-based backbones—RemoteCLIP [33], GeoRSCLIP [59], and original CLIP [44]—paired with bilinear upsampling and RIFI-Up. As shown in Tab. 4, Free-CD achieves competitive OVCD performance across all backbone-upsampler pairings, even when bypassing the DINO-based feature extractor. This demonstrates the framework's robust adaptability and its capacity to rapidly operationalize existing OVSS models as practical OVCD solutions with minimal integration overhead.

Table 3: Ablation of Upsampling Strategies. Bilinear: bilinear interpolation; JBU: joint bilateral upsampling; Conv: transposed convolution; FI: feature inversion; RI: resolution invariant. The best and second-best results in each column are highlighted.

	Upsampler	FI	RI	SECOND		HiUCD		HRSCD		JL1-SCD	
				Fscd	mIoU	Fscd	mIoU	Fscd	mIoU	Fscd	mIoU
Training-free	Bilinear	✗	-	21.11	56.32	28.57	52.25	8.23	57.22	15.14	49.64
	JBU	✗	-	21.47	56.43	28.73	53.23	8.24	57.27	15.51	50.44
Pretrained	Conv	✗	-	18.72	54.07	28.97	53.85	7.89	56.90	15.18	50.91
	JBU	✗	-	21.53	57.91	29.56	54.40	8.29	57.36	15.77	51.54
	JBU	✓	✗	22.18	58.87	29.74	54.54	8.18	57.30	15.62	51.43
	JBU	✓	✓	23.19	59.84	30.00	54.99	8.37	57.73	16.15	52.37

Table 4: Adaptability of Free-CD to different CLIP-based backbones and upsamplers. The best and second-best results in each column are highlighted.

Pretrained	Upsampler	SECOND		HiUCD		HRSCD		JL1-SCD	
		Fscd	mIoU	Fscd	mIoU	Fscd	mIoU	Fscd	mIoU
CLIP	Bilinear	16.71	46.30	18.91	46.79	4.81	43.39	4.88	37.71
	RIFI-Up	17.27	48.33	19.19	48.91	4.03	44.76	6.27	44.06
RemoteCLIP	Bilinear	17.04	47.53	14.69	43.68	4.04	39.68	4.90	40.06
	RIFI-Up	18.18	49.16	16.42	46.14	3.57	41.23	11.65	46.32
GeoRSCLIP	Bilinear	17.52	47.62	18.72	45.66	4.78	42.69	4.75	41.94
	RIFI-Up	19.42	49.81	18.60	48.16	4.19	43.13	12.05	48.71

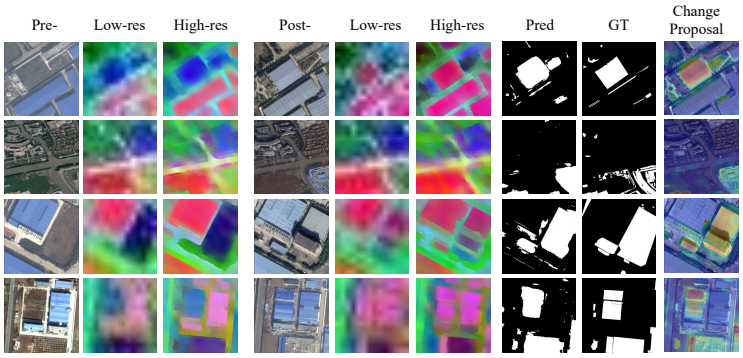

Fig. 6: Feature visualization results. **Pre-** and **Post-** denote the bi-temporal input images. **Low-res** and **High-res** represent the original low-resolution features and the upsampled high-resolution features generated by RIFI-Up, respectively. **Pred** indicates the predicted change mask inferred by the model, while **GT** corresponds to the ground truth mask. The **Change Proposal** is visualized as a heatmap highlighting potential change regions.

6.5 Feature Visualization

To qualitatively assess the semantic fidelity of our proposed method, we employ Principal Component Analysis (PCA) to visualize the feature representations. As illustrated in Fig. 6, RIFI-Up demonstrates superior high-fidelity upsampling capabilities. Consequently, the derived change proposals and predicted masks display significantly reduced artifacts, closely aligning with the ground truth.

7 Conclusion

In this paper, we presented Free-CD, a unified, training-free framework for open-vocabulary change detection that decouples change localization from semantic classification. By integrating RIFI-Up for resolution-invariant feature recovery and Bayesian Probability Correction to mitigate semantic ambiguity, Free-CD achieves superior spatial fidelity and robustness without task-specific fine-tuning. Extensive experiments across four benchmarks confirm that our method consistently outperforms state-of-the-art approaches in accuracy and generalization. The framework offers a versatile pipeline for evolving, practical open-vocabulary change analysis.

Acknowledgements

The work was supported by the National Natural Science Foundation of China under Grants 62271018, 62125102, and 623B2013, and the Beijing Natural Science Foundation under Grant JL23005.

References

1. Augustine, M.T.: A survey on universal approximation theorems. arXiv preprint arXiv:2407.12895 (2024)
2. Bai, B., Fu, W., Lu, T., Li, S.: Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021)
3. Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. pp. 207–210. IEEE (2022)
4. Bovolo, F., Bruzzone, L.: A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. IEEE Transactions on Geoscience and Remote Sensing **45**(1), 218–236 (2007)
5. Cao, Q., Chen, Y., Ma, C., Yang, X.: Open-vocabulary high-resolution remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing (2025)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)

7. Celik, T.: Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE geoscience and remote sensing letters **6**(4), 772–776 (2009)
8. Chen, H., Song, J., Han, C., Xia, J., Yokoya, N.: Changemamba: Remote sensing change detection with spatio-temporal state space model. arXiv preprint arXiv:2404.03425 (2024)
9. Chen, K., Chen, B., Liu, C., Li, W., Zou, Z., Shi, Z.: Rsmamba: Remote sensing image classification with state space model. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024)
10. Chen, K., Liu, C., Li, W., Liu, Z., Chen, H., Zhang, H., Zou, Z., Shi, Z.: Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection. arXiv preprint arXiv:2312.16202 (2023)
11. Chen, K., Liu, C., Chen, B., Zhang, J., Zou, Z., Shi, Z.: Rsrefseg 2: Decoupling referring remote sensing image segmentation with foundation models. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–20 (2025)
12. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024)
13. Chen, K., Liu, C., Li, W., Chen, B., Zou, Z., Lu, S., Shi, Z.: Remote sensing image change captioning: A comprehensive survey. Available at SSRN 6836181 (2026)
14. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
15. Daudt, R.C., Le Saux, B., Boulch, A.: Fully convolutional siamese networks for change detection. In: 2018 25th IEEE international conference on image processing (ICIP). pp. 4063–4067. IEEE (2018)
16. Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y.: Multitask learning for large-scale semantic change detection. Computer Vision and Image Understanding **187**, 102783 (2019)
17. Ding, L., Zhang, J., Guo, H., Zhang, K., Liu, B., Bruzzone, L.: Joint spatio-temporal modeling for semantic change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (2024)
18. Du, B., Ru, L., Wu, C., Zhang, L.: Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **57**(12), 9976–9992 (2019)
19. El Amin, M.A.M., Liu, Q., Wang, Y.: Convolutional neural network features based change detection in satellite images. In: First international workshop on pattern recognition. vol. 10011, p. 181. SPIE (2016)
20. Fang, S., Li, K., Shao, J., Li, Z.: Snunet-cd: A densely connected siamese network for change detection of vhr images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021)
21. Fu, S., Hamilton, M., Brandt, L.E., Feldmann, A., Zhang, Z., Freeman, W.: Featup: A model-agnostic framework for features at any resolution. In: International Conference on Learning Representations. vol. 2024, pp. 45324–45350 (2024)
22. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
23. Guo, H., Liu, C., Zhang, H., Chen, B., Zou, Z., Shi, Z.: Taco: Capturing spatio-temporal semantic consistency in remote sensing change detection. arXiv preprint arXiv:2511.20306 (2025)

24. Guo, Q., Wang, R., Huang, R., Wan, R., Sun, S., Zhang, Y.: Idet: Iterative difference-enhanced transformers for high-quality change detection. IEEE Transactions on Emerging Topics in Computational Intelligence (2025)
25. Han, C., Zhong, Y., Li, D., Han, K., Ma, L.: Open-vocabulary semantic segmentation with decoupled one-pass network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1086–1096 (2023)
26. Ji, S., Wei, S., Lu, M.: 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 (2018)
27. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
28. Li, B., Dong, H., Zhang, D., Zhao, Z., Gao, J., Li, X.: Exploring efficient open-vocabulary segmentation in the remote sensing. arXiv preprint arXiv:2509.12040 (2025)
29. Li, K., Cao, X., Deng, Y., Pang, C., Xin, Z., Qiao, H., Gong, T., Meng, D., Wang, Z.: Dynamicearth: How far are we from open-vocabulary change detection? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6279–6287 (2026)
30. Li, K., Liu, R., Cao, X., Bai, X., Zhou, F., Meng, D., Wang, Z.: Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10545–10556 (2025)
31. Li, M., Ming, D., Xu, L., Dong, D., Zhang, Y.: Sfearnet: A network combining semantic flow and edge-aware refinement for highly efficient remote sensing image change detection. IEEE Transactions on Geoscience and Remote Sensing **63**, 3545906 (2025)
32. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
33. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024). https://doi.org/10.1109/TGRS.2024.3390838
34. Liu, Y., Bai, S., Li, G., Wang, Y., Tang, Y.: Open-vocabulary segmentation with semantic-assisted calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3491–3500 (2024)
35. Long, X., Zhuang, W., Xia, M., Hu, K., Lin, H.: Sasiamnet: Self-adaptive siamese network for change detection of remote sensing image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 1021–1034 (2023)
36. Long, Y., Xia, G.S., Li, S., Yang, W., Yang, M.Y., Zhu, X.X., Zhang, L., Li, D.: On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. IEEE Journal of selected topics in applied earth observations and remote sensing **14**, 4205–4230 (2021)
37. Nikolaou, G., Mencattini, T., Crisostomi, D., Santilli, A., Panagakis, Y., Rodolà, E.: Language models are injective and hence invertible. arXiv preprint arXiv:2510.15511 (2025)

38. Niu, Y., Guo, H., Lu, J., Ding, L., Yu, D.: Smnet: symmetric multi-task network for semantic change detection in remote sensing images based on cnn and transformer. Remote Sensing **15**(4), 949 (2023)
39. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
40. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)
41. Peng, D., Bruzzone, L., Zhang, Y., Guan, H., He, P.: Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. International Journal of Applied Earth Observation and Geoinformation **103**, 102465 (2021)
42. Peng, X., Zhong, R., Li, Z., Li, Q.: Optical remote sensing image change detection based on attention mechanism and image difference. IEEE Transactions on Geoscience and Remote Sensing **59**(9), 7296–7307 (2020)
43. Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., et al.: Freeseg: Unified, universal and open-vocabulary image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19446–19455 (2023)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision (2021)
45. Saha, S., Bovolo, F., Bruzzone, L.: Unsupervised deep change vector analysis for multiple-change detection in vhr images. IEEE Transactions on Geoscience and Remote Sensing **57**(6), 3677–3693 (2019)
46. Shen, Y., Fu, C., Chen, P., Zhang, M., Li, K., Sun, X., Wu, Y., Lin, S., Ji, R.: Aligning and prompting everything all at once for universal visual perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13193–13203 (2024)
47. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
48. Tan, X., Chen, G., Wang, T., Wang, J., Zhang, X.: Segment change model (scm) for unsupervised change detection in vhr remote sensing images: a case study of buildings. In: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. pp. 8577–8580. IEEE (2024)
49. Tang, X., Zhang, H., Mou, L., Liu, F., Zhang, X., Zhu, X.X., Jiao, L.: An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2021)
50. Tian, S., Ma, A., Zheng, Z., Tan, X., Zhong, Y.: Learning temporal consistency for high spatial resolution remote sensing imagery semantic change detection. IEEE Transactions on Geoscience and Remote Sensing (2025)
51. Tian, S., Ma, A., Zheng, Z., Zhong, Y.: Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery. arXiv preprint arXiv:2011.03247 (2020)
52. Xiang, Y., Tian, X., Xu, Y., Guan, X., Chen, Z.: Egmt-cd: Edge-guided multimodal transformers change detection from satellite and aerial images. Remote Sensing **16**(1), 86 (2023)

53. Xie, B., Cao, J., Xie, J., Khan, F.S., Pang, Y.: Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3426–3436 (2024)
54. Xiong, Z., Wang, Y., Yu, W., Stewart, A.J., Zhao, J., Lehmann, N., Dujardin, T., Yuan, Z., Ghamisi, P., Zhu, X.X.: Dofa-clip: Multimodal vision-language foundation models for earth observation. arXiv preprint arXiv:2503.06312 (2025)
55. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
56. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
57. Yang, K., Xia, G.S., Liu, Z., Du, B., Yang, W., Pelillo, M., Zhang, L.: Semantic change detection with asymmetric siamese networks. arXiv preprint arXiv:2010.05687 (2020)
58. Ye, C., Zhuge, Y., Zhang, P.: Towards open-vocabulary remote sensing image semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9436–9444 (2025)
59. Zhang, Z., Zhao, T., Guo, Y., Yin, J.: Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2024)
60. Zheng, Z., Zhong, Y., Zhang, L., Ermon, S.: Segment any change. Advances in Neural Information Processing Systems **37**, 81204–81224 (2024)
61. Zhu, S., Song, Y., Zhang, Y., Zhang, Y.: Ecfnet: A siamese network with fewer fps and fewer fns for change detection of remote-sensing images. IEEE Geoscience and Remote Sensing Letters **20**, 1–5 (2023)
62. Zhu, X.X., Xiong, Z., Wang, Y., Stewart, A.J., Heidler, K., Wang, Y., Yuan, Z., Dujardin, T., Xu, Q., Shi, Y.: On the foundations of earth foundation models. Communications Earth & Environment (2026)