DynamicEarth: How Far Are We from Open-Vocabulary Change Detection?

Kaiyu Li[1], Xiangyong Cao[†2,8], Yupeng Deng[3], Chao Pang[4], Zepeng Xin[5], Hui Qiao[6], Tieliang Gong[2], Deyu Meng [7,8], Zhi Wang[1]

[1]School of Software Engineering, Xi'an Jiaotong University
[2]School of Computer Science and Technology, Xi'an Jiaotong University
[3]Aerospace Information Research Institute, Chinese Academy of Sciences
[4]School of Computer Science, Wuhan University
[5]College of Artificial Intelligence, Xi'an Jiaotong University
[6]China Telecom Shaanxi Branch
[7]School of Mathematics and Statistics, Xi'an Jiaotong University
[8]Ministry of Education Key Laboratory of Intelligent Networks and Network Security

Abstract

Monitoring Earth's evolving land covers requires methods capable of detecting changes across a wide range of categories and contexts. Existing change detection methods are hindered by their dependency on predefined classes, reducing their effectiveness in open-world applications. To address this issue, we introduce open-vocabulary change detection (OVCD), a novel task that bridges vision and language to detect changes across any category. Considering the lack of high-quality data and annotation, we propose two training-free frameworks, M-C-I and I-M-C, which leverage and integrate off-the-shelf foundation models for the OVCD task. The insight behind the M-C-I framework is to discover all potential changes and then classify these changes, while the insight of I-M-C framework is to identify all targets of interest and then determine whether their states have changed. Based on these two frameworks, we instantiate to obtain several methods, *e.g.*, SAM-DINOv2-SegEarth-OV, Grounding-DINO-SAM2-DINO, *etc*. Extensive evaluations on 4 benchmark datasets demonstrate the superior generalization and robustness of our OVCD methods over existing supervised and unsupervised methods. To support continued exploration, we release DynamicEarth, a dedicated codebase designed to advance research and application of OVCD.

Code — https://github.com/likyoo/DynamicEarth

Introduction

The Earth is a dynamic system in a state of perpetual evolution. Observing this vibrant planet allows us to deepen our understanding of intricate phenomena, including human activity, geographic evolution, and climate change. This process requires the continuous monitoring of land use and land cover types and provides insights into where changes have occurred and the particulars of those changes. The most pertinent task in this context is change detection (CD), which specifically involves analyzing bi-temporal or multi-temporal satellite and aerial images to determine what changes are occurring where.

[†]Corresponding author (caoxiangyong@mail.xjtu.edu.cn)

As a higher-level computer vision task beyond segmentation and detection, CD technology has made significant strides with the support of basic vision techniques. Contemporary supervised learning approaches typically process bi-temporal images through Siamese networks to generate change masks or bounding box predictions (Fang, Li, and Li 2023). However, such supervised approaches demonstrate constrained generalizability even when handling identical object categories (*e.g.*, buildings) across disparate imaging conditions, particularly variations in camera sensors and ground sample distances (GSD). Recent advancements in foundation models have demonstrated enhanced generalization across supervised (Li, Cao, and Meng 2024), semi-supervised (Li et al. 2024a), and unsupervised (Tan et al. 2023; Zheng et al. 2024) CD frameworks by capitalizing on their inherent general knowledge. In particular, AnyChange (Zheng et al. 2024) leverages the mask proposal and feature matching of the segment anything model (SAM) (Kirillov et al. 2023) to construct a universal unsupervised CD model without any post-training. Nevertheless, its applicability remains confined to binary CD due to the generation of class-agnostic change masks. This critical limitation implies that while the method successfully localizes where changes occur, it fails to characterize what semantic categories undergo changes. Motivated by this, we introduce the **Open-Vocabulary Change Detection (OVCD)** task, which aims to detect changes for arbitrary user-specified categories of interest. Furthermore, we also explore how to build an OVCD system and how far current methodologies are from achieving OVCD.

Different from open-vocabulary segmentation or detection of single-temporal images (Zhu and Chen 2024), OVCD involves the identification and comparison of bi- or multi-temporal images. In addition, Prior studies (Li et al. 2024a; Zheng et al. 2024) emphasize that instance-level comparisons are generally superior to simple pixel-level comparisons in CD, effectively mitigating pixel-level pseudo changes that may arise from changes in lighting, season, viewpoint, *etc*. This empirical evidence establishes instance-level object/mask proposal as an indispensable component of OVCD systems. Accordingly, we posit that a comprehen-

*"The intelligent earth observation system of the future can tackle the "**4W**" queries, i.e., when, where, what object, and what change has occurred."* -- *Deren Li*

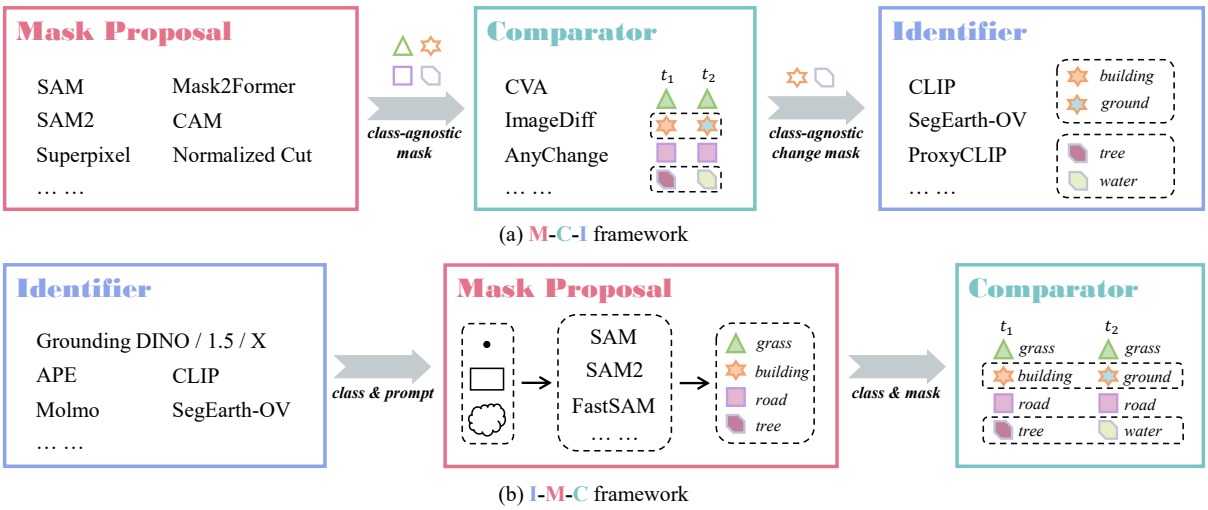

Figure 1: The two OVCD frameworks proposed in this paper. (a) M-C-I: discover all class-agnostic masks, determine whether the mask region has changed, and identify the change class. (b) I-M-C: identify all targets of interest, convert to mask format, and compare whether the target state has changed.

sive OVCD framework should consist of three components: an identifier, a comparator, and the proposal of instance-level objects/masks. Given the scarcity of high-quality annotations for CD, we do not suggest training a specialized vision-language model (VLM) for CD from scratch, but rather reusing off-the-shelf general VLMs.

Based on the above, in this paper, we propose two training-free universal frameworks for OVCD, M-C-I and I-M-C, as shown in Figure 1. **(1)** M-C-I framework first generates class-agnostic masks using mask proposal methods, *e.g.*, SAM (Kirillov et al. 2023), Mask2Former (Cheng et al. 2022), *etc*. It then compares the region corresponding to each mask in the bi-temporal images or features to determine whether a change has occurred. Finally, the change region is fed into the open-vocabulary classifier (*e.g.*, CLIP (Radford et al. 2021), SegEarth-OV (Li et al. 2025a)) to identify the change category. **(2)** I-M-C framework is inspired by the post-classification comparison (PCC) method (Howarth and Wickware 1981). It first leverages perceptual foundation models (e.g., Grounding DINO (Liu et al. 2025), APE (Shen et al. 2024), Molmo (Deitke et al. 2024)) for text-guided object localization, generating preliminary geometric cues (bounding boxes, coarse masks, or key points). These cues serve as visual prompts to refine segmentation through mask proposal methods, followed by spatial consistency analysis of bi-temporal masks to detect changes.

To summarize, the contributions of this work include:

- A new task, OVCD, is introduced, which unlocks language-guided CD and allows detection of changes in any category.
- Inspired by the *"4W"* assertion (see Figure 1) and considering the constraints of limited annotated data, we propose two **training-free** universal frameworks for OVCD by fully reusing the off-the-shelf foundation models.
- Extensive evaluations on 4 datasets highlight the generalization and robustness of our method, significantly surpassing existing unsupervised and supervised methods.

Related Work

Vision-language Model. VLMs aim to bridge the gap between visual and textual modalities, enabling systems to understand and generate multi-modal content. VLMs like CLIP (Radford et al. 2021) align visual and text representations. Recent models (Liu et al. 2024), inspired by large language model (LLM) (Touvron et al. 2023), have evolved into general perceptual systems. Grounding DINO (Liu et al. 2025) marries DINO (Zhang et al. 2022) with grounded pre-training for open-set object detection (Li et al. 2025b, 2024d). Then, the subsequent APE (Shen et al. 2024) and DINO-X (Ren et al. 2024b) allow fine-grained perception *e.g.* segmentation or keypoints. Further, the advanced Florence-2 (Xiao et al. 2024) takes textual prompts as task instructions and generates desirable results in text form, including captioning, object detection and segmentation.

Segment Anything Model. SAM (Kirillov et al. 2023) pioneers a new segmentation paradigm, utilizing prompt-based learning to enable segmentation using points or boxes as inputs. Unlike traditional methods that rely on domain-specific training datasets, SAM leverages vast amounts of pre-trained data to achieve high generalizability. Based on this, HQ-SAM (Ke et al. 2024) introduces HQ-Token for high-quality mask prediction. Recent advancements include models such as FastSAM (Zhao et al. 2023), which focus on improving efficiency, scalability, and real-time performance.

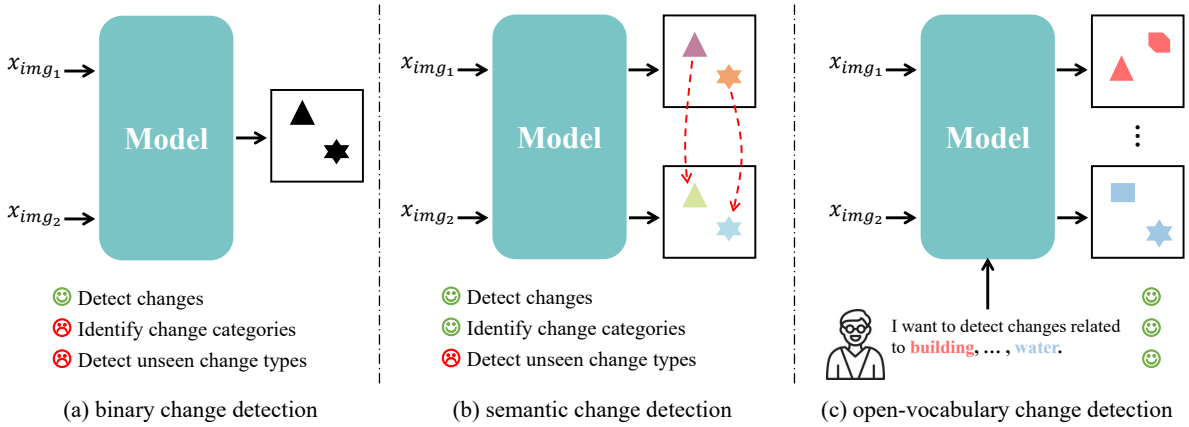

Figure 2: Different CD tasks: (a) Binary CD aims at discovering all (interested) changes and generating a binary mask; (b) Semantic CD further identifies the category of changes. However, both can only be trained and evaluated on data with predefined categories; (c) Our proposed OVCD can detect changes in any category according to the user's requirements.

Most recently, SAM 2 (Ravi et al. 2024) has been proposed to further extend to video, while keeping the original ability to segment everything in images. In OVCD, SAMs can be used either as an initial step to propose all candidate regions or as a post-refiner of the identification results.

Open-vocabulary Semantic Segmentation. OVSS extends semantic segmentation to recognize and segment unseen categories at inference time. Most OVSS methods build on top of VLMs. Early OVSS methods are inspired by vision-language contrastive learning and try to train the CLIP variant with pixel-level perception (Liang et al. 2023; Cho et al. 2024). Due to the domain difference between remote sensing images and natural images, Cao *et.al.* (Cao et al. 2024) proposed a CLIP fine-tuning based method and trained it on some remote sensing segmentation data. Similarly, Ye *et.al.* (Ye, Zhuge, and Zhang 2025) used more remote sensing data (consisting of public datasets) to train both CLIP and a specialist backbone. Different from the above, Li *et.al.* (Li et al. 2025a) found that feature resolution is the key factor constraining OVSS of remote sensing images, and proposed a training-free OVSS method, SegEarth-OV, which even outperforms the training-based method. In this paper, we believe that identification is an important component of OVCD, and thus to some extent, a feasible OVSS method is the prerequisite for OVCD.

Binary CD. Binary CD involves identifying regions of change between two temporally separated images. Traditional methods *e.g.*, change vector analysis (CVA) (Bovolo and Bruzzone 2006) relied on pixel-wise comparisons, while more recent methods leverage deep learning for enhanced feature extraction and robustness. In the foundation model era, BAN (Li, Cao, and Meng 2024) introduces VLMs to binary CD and provides a supervised parameter efficient fine-tuning solution. Subsequently, Li *et.al.* (Li et al. 2024a) used the pseudo labels generated by VLMs as additional supervised signals to improve the performance of semi-supervised CD. Zheng *et.al.* (Zheng et al. 2024) proposed AnyChange for unsupervised binary CD, which yields high-recall class-agnostic change masks. However, in practical cases, the extraction of changes of interest is essential, and our proposed M-C-I framework can be regarded as a generalization of AnyChange, *i.e.*, it can segment any change of interest.

Semantic CD. Semantic CD builds on binary CD by not only identifying changes but also classifying them into semantic categories (Li et al. 2024b; Ding et al. 2025). As mentioned above, compared to binary CD, semantic CD is more in line with practical applications. Users often only need to detect changes in some specific categories, *e.g.*, building CD for urban expansion analysis (Ji, Wei, and Lu 2018; Chen and Shi 2020), cropland CD for agricultural protection, landslide CD for disaster monitoring, *etc*. Typical semantic CD methods follow a triple decoder architecture, *i.e.*, one difference branch for binary CD and two semantic branches for bi-temporal class discrimination. In this paper, we simplify the multi-class semantic CD to single-class semantic CD, which achieves the same objective while avoiding the complexity of the model structure, and allows the extraction of changes in any class using only off-the-shelf single-temporal models.

Problem Definition

OVCD aims to localize and identify changes between two temporally separated images x_{img_1} and x_{img_2} of the same scene, where the categories of changes are not predefined and can be described by arbitrary textual or semantic labels x_{text}. OVCD extends traditional CD by introducing the ability to generalize beyond a fixed set of predefined change classes, enabling the detection and interpretation of novel changes using textual guidance or contextual understanding, as shown in Figure 2. The task shares a similar formulation with the OVSS task (Li et al. 2025a), but is far more challenging. The key challenge lies in:

- **Bi- or Multi-temporal Image Input.** Beyond OVSS, OVCD involves comparisons between image pairs. This introduces several questions, *e.g.*, when should comparisons be made, before or after identification? What level

of comparison should be performed, pixel level or instance level? How to mitigate error accumulation due to multiple steps? And so on.

- **Cross-domain from Natural to Remote Sensing Images.** Remote sensing images are generally acquired by sensors from satellites or aircraft, and they present a bird's eye view, as different from natural images which mostly present a horizontal view. These two views bring completely different surface features when facing the same object, *e.g.*, for a building, it may be characterized by a window in the horizontal view and a roof in the overhead view. This difference leads to the fact that most of the models trained in natural images cannot be directly applied to remote sensing images.
- **Other Challenges.** There are also some challenges due to the characteristics of remote sensing images, *e.g.*, OVCD of non-RGB images (*i.e.*, multispectral, hyperspectral, SAR images, *etc*.), object scale spanning and small object issues in overhead views.

Method

Following the experience of OVSS, there are two basic paths to achieve OVCD: training-based and training-free paradigms. Training-based methods generally train on some basic classes and then generalize to novel classes by leveraging the cross-modal and zero-shot capabilities of VLMs. However, the current volume of CD data, especially semantic CD data, is limited and dispersed under various GSDs (Peng et al. 2025). In addition, the image quality and annotation quality are worrisome (Tian et al. 2020). These factors make it difficult to build a complete training-based OVCD pipeline. On the other hand, some recent studies have demonstrated that foundation models trained on web data or natural image data also exhibit generalization capabilities in remote sensing scenarios (Shen et al. 2024; Li et al. 2024a, 2025a; Zheng et al. 2024). For instance, SegEarth-OV (Li et al. 2025a) indicates that standard CLIP (Radford et al. 2021) can be directly used for semantic segmentation of remote sensing images without any post-training.

Therefore, in this paper, we focus on ***how to construct training-free OVCD methods using off-the-shelf foundation models***. Based on the definitions of CD and OVCD, we find two strategies to achieve "segment any change of interest": (1) discover all potential changes, and then classify these changes; (2) identify all targets of interest, and then determine whether their states have changed between two images. Therefore, a **comparator** and an open-set **identifier** are necessary to implement these two strategies. In addition, in practice, the **mask proposal** is also needed. For the first strategy, the mask proposal can mitigate pixel-level pseudo changes that may arise from changes in lighting, season, viewpoint, *etc*. For the second strategy, the mask proposal can convert the bounding boxes and points in the identification results into uniform masks, aligning more effectively with the task requirements. According to the above two strategies, we propose two training-free OVCD frameworks, M-C-I and I-M-C, as shown in Figure 1.

M-C-I Framework

The M-C-I framework pipeline begins with discovering all class-agnostic masks. Following this, it compares whether the bi-temporal features corresponding to these masks have changed. Finally, it filters the changes of interest. Based on this, this framework is divided into three sequential components: **Mask Proposal**, **Comparator**, and **Identifier**.

Mask Proposal component in the M-C-I framework is designed to divide the image into several partitions $\mathcal{M}_t = \{\mathbf{m}_{t,i}\}_{i\in[1,2,\dots,N_t]}$, where t indicates the temporal index, and expects each partition to contain an instance or a single category of pixels. In fact, before the deep learning era, some traditional image segmentation algorithms had this capability, *e.g.*, Normalized Cuts (Shi and Malik 2000), SuperPixel (Achanta et al. 2012), Ncuts (Ren and Malik 2003), *etc*. These methods only utilize texture properties to segment the image and thus have limited performance on complex and low-resolution images. A better alternative is SAMs, which are trained on a large number of high-quality labeled images (*e.g.*, SAM (Kirillov et al. 2023) uses 11M images with 1.1B masks, SAM 2 (Ravi et al. 2024) uses 1.4K videos with 16K masklets), and thus offer strong generalization and fine segmentation (Chen et al. 2024; Ren et al. 2024a). In this framework, SAMs are required to enable the automatic mask generation mode to generate all candidate masks. Since the changed objects may appear in either temporal image, we concatenate the bi-temporal candidate mask sets as an overall candidate mask set. Considering that the bi-temporal images are captured from the same region and contain numerous unchanged objects, we use non-maximum suppression (NMS) to remove duplicate masks. Specifically, to enhance efficiency, we use the outer bounding boxes and the IoU predictions of candidate masks as inputs to the NMS. Consequently, the final output of the Mask Proposal can be denoted as $\mathcal{M} = \mathit{NMS}(\mathcal{M}_1 \cup \mathcal{M}_2)$.

Comparator is used to discriminate whether the mask region has changed or not, and its purpose is the same as the binary CD task. Therefore, some traditional CD methods can be used as comparators, *e.g.*, CVA (Malila 1980), ratioing method (Lu et al. 2004), differencing method (Mahmoudzadeh 2007), *etc*. In our implementation, following AnyChange (Zheng et al. 2024), we use a latent matching method that use negative cosine similarity as the change score for bi-temporal latent features in the mask region (higher score indicates more significant change). But unlike AnyChange, we suggest to use DINO/DINOv2 (Zhang et al. 2022; Oquab et al. 2023) to extract features, considering the natural advantage of DINOs, *i.e.*, they are trained using contrastive learning, where the feature distances of the same objects are pulled closer together, and vice versa. Specifically, the change score can be formulated as:

$$\mathcal{D}(\mathbf{m}) = -\frac{\mathbf{z_1}[\mathbf{m}]}{\|\mathbf{z_1}[\mathbf{m}]\|_2} \cdot \frac{\mathbf{z_2}[\mathbf{m}]}{\|\mathbf{z_2}[\mathbf{m}]\|_2}, \tag{1}$$

where the vector $\mathbf{z_t}[\mathbf{m}]$ indicates the average of the feature maps from x_{img_t} extracted using DINOs at the indexes corresponding to the mask $\mathbf{m}$. $\mathbf{m} \in \mathcal{M}$ and $t \in \{1, 2\}$. The masks with change scores higher than the threshold β will

be discriminated as change masks and the rest are dropped.

Identifier is used to filter specific categories of changes from all change masks. Since the categories of interest to the user may be varied, it is essential to employ VLMs with open-set identification capabilities, *e.g.*, CLIP and its derived models, as the identifier. A straightforward method is to crop out the image regions corresponding to each change mask and extract their global `[CLS]` tokens using CLIP, and then compute the similarity between these global `[CLS]` tokens and the text embedding of the category of interest to obtain the final identification result. However, this method requires feeding the image patches corresponding to each change mask into CLIP's image encoder, resulting in high computational cost. Inspired by SegEarth-OV (Li et al. 2025a), which demonstrates that patch tokens, *i.e.*, feature map, generated by CLIP can also be used for collaborative inference with text embeddings in remote sensing scenarios. In our implementation, we extract the full-image features only once for x_{img_1} and x_{img_2}, and then use change masks to crop the corresponding regions in the feature map, and finally calculate the average vector as the image representation of the masks (*a.k.a.* masked average pooling), which is used in conjunction with the text embedding for inference.

I-M-C Framework

The I-M-C framework is inspired by the PCC method. It initially identifies target instances of interest and locates them in the form of boxes, points, or masks. Subsequently, it converts these instances into a uniform mask format. Finally, the states of the bi-temporal targets at the corresponding positions are compared to determine whether any changes have occurred. Thus, the I-M-C framework can be composed of three sequential components: **Identifier**, **Mask Proposal** and **Comparator**.

Identifier here is different from that in M-C-I, as it must not only identify arbitrary categories of objects but also initially discover all targets. To achieve this, some open-vocabulary detection, visual grounding, or even multi-modal large language models (MLLMs) can be selected as identifiers. For example, we process the bi-temporal images separately, using Grounding DINOs (Liu et al. 2025; Zhao et al. 2024) to generate bounding boxes, Molmo (Deitke et al. 2024) to generate points, or APE (Shen et al. 2024) to generate masks directly for objects of interest.

Mask Proposal in the I-M-C framework is used to convert the bounding box, point, or coarse mask generated by the identifier into a uniform finer mask, as the OVCD task ultimately requires pixel-level output. The demand of this component directs to interactive segmentation models *e.g.* SAM, SAM 2, FastSAM, *etc*. In our implementation, we feed the bi-temporal images along with their respective identification results into the interactive segmentation model. Thus, all instance masks and their categories for each of the bi-temporal images are obtained.

Comparator in the I-M-C framework has a similar capability as the comparator in the M-C-I framework, *i.e.*, to determine whether the candidate mask region has changed. However, unlike the former, in the I-M-C framework, the categories of the candidate masks are known. Therefore, a simple geometry-based comparison method is logically sufficient. Following (Li et al. 2024a), we use an IoU-aware method. Specifically, an instance mask is considered as unchanged if the sum of its IoUs with all the remaining instance masks of the same category is higher than a predefined threshold; otherwise, it is regarded as changed. However, in practice, due to the limited capability of the identifier, it is common that for the identical object, it can be detected in one image but missed in another, resulting in pseudo change. To alleviate this issue, we additionally use the latent matching method in M-C-I, where each masked region is individually compared at the feature level. Finally, only the change masks confirmed by both comparison methods are kept.

Method	LEVIR-CD		WHU-CD		S2Looking	
	IoU^c	F_1^c	IoU^c	F_1^c	IoU^c	F_1^c
PCA-KM	4.8	9.1	5.4	10.2	-	-
CNN-CD	7.0	13.1	4.9	9.4	-	-
DSFA	4.3	8.2	4.1	7.8	-	-
DCVA	7.6	14.1	10.9	19.6	-	-
GMCD	6.1	11.6	10.9	19.7	-	-
CVA	-	12.2	-	-	-	5.8
DINOv2+CVA	-	17.3	-	-	-	4.3
AnyChange-H	-	23.0	-	-	-	6.4
SCM	18.8	31.7	18.6	31.3	-	-
M-C-I:						
SAM - DINO - SegOV	33.0	49.7	36.7	53.7	22.5	36.7
SAM - DINOv2 - SegOV	36.6	53.6	40.6	57.7	**23.9**	**38.5**
SAM2 - DINOv2 - SegOV	33.8	50.5	40.9	58.1	23.1	37.6
I-M-C:						
MMGD - SAM2 - DINO	15.6	27.0	11.0	19.8	2.3	4.5
APE - / - DINO	**53.5**	**69.7**	56.8	72.5	10.1	18.4
APE - / - DINOv2	50.0	66.7	**61.1**	**75.8**	5.3	10.1

Table 1: OVCD quantitative comparison on building CD datasets. "SegOV" denotes SegEarth-OV (Li et al. 2025a), "MMGD" denotes MM-Grounding-DINO (Zhao et al. 2024), and "-" denotes data missing.

Experiment

Dataset

To fully assess the proposed M-C-I and I-M-C frameworks, three building CD and one land cover CD datasets are selected. Since both proposed frameworks are training-free, we mainly focus on their test/validation sets.

LEVIR-CD dataset (Chen and Shi 2020) is collected from Google Earth with a spatial resolution of 0.5m/pixels. Its test set contains 128 pairs of 1,024×1,024 images. When processing with the I-M-C framework, considering the small target issue in remote sensing images, we crop the original images into non-overlapping 256×256 image patches.

WHU-CD dataset (Ji, Wei, and Lu 2018) contains 7,434 aerial image pairs with a size of 256×256 and 744 for testing. Its spatial resolution is 0.2m/pixels.

S2Looking dataset (Shen et al. 2021) contains 1,000 1,024×1,024 test data with a spatial resolution of 0.5∼0.8m/pixel. It is also pre-cropped into 256×256 patches for I-M-C framework.

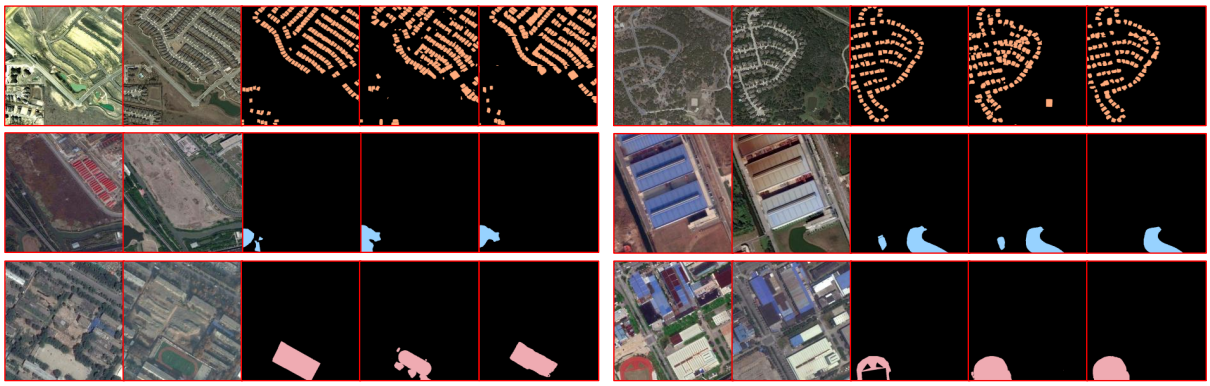

Figure 3: Open-vocabulary CD examples. In each group: x_{img_1}, x_{img_2}, ground truth, the result of an M-C-I method and the result of an I-M-C method. Color rendering: orange: "building", blue: "water" pink: "playground".

Method	Building		Tree		Water		Low vegetation		N.v.g surface		Playground	
	IoU^c	F_1^c	IoU^c	F_1^c	IoU^c	F_1^c	IoU^c	F_1^c	IoU^c	F_1^c	IoU^c	F_1^c
M-C-I:												
SAM - DINO - SegOV	34.1	50.8	16.5	28.3	13.4	23.6	24.0	38.7	22.5	36.7	16.0	27.6
SAM - DINOv2 - SegOV	**38.1**	**55.2**	**20.3**	**33.8**	**14.3**	**25.1**	**24.1**	**38.9**	**26.2**	**41.6**	**20.0**	**33.3**
SAM2 - DINOv2 - SegOV	36.6	53.5	18.2	30.8	13.8	24.3	22.1	36.2	19.2	32.3	17.1	29.2
I-M-C:												
MMGD - SAM2 - DINO	9.5	17.4	7.0	13.1	1.2	2.3	5.2	9.8	1.0	2.0	-	-
APE - / - DINO	26.5	42.0	13.5	23.8	9.8	17.9	-	-	-	-	16.5	28.3
APE - / - DINOv2	28.1	43.9	14.1	24.8	12.2	21.7	1.4	2.7	-	-	16.0	27.6

Table 2: OVCD quantitative comparison on SECOND dataset. "-" denotes that the score is close to 0.

SECOND dataset (Yang et al. 2021) focus on 6 main land-cover classes, *i.e.*, non-vegetated ground surface (N.v.g surface), tree, low vegetation, water, building, and playground. It contains 4,662 pairs of aerial images and 593 for testing.

Implementation Details

Codebase. Since this paper presents the OVCD task and framework for the first time, we develop a PyTorch-based codebase, DynamicEarth, for the reimplementation of the proposed methods and subsequent research. The core code of DynamicEarth is organized to match both frameworks, *i.e.*, it includes identifiers, comparators, and mask proposals. In addition, DynamicEarth includes both evaluation scripts for each dataset and demo scripts for each method with detailed comments and standardized code. The different scripts are isolated, and we believe this design is user-friendly to understand, debug, and contribute code.

Setup. All methods proposed in this paper can be run on a single 4090 GPU. The setting of the hyperparameter β fluctuates across methods or categories, and we endeavor to find its optimal value. For text prompts, due to different preferences in various VLMs, we will discuss this below.

Evaluation. For all datasets, we calculate the IoU and F1 score corresponding to each class separately, which are denoted as IoU^c and F_1^c.

Compared Method. Since both frameworks proposed in this paper are training-free, we select some unsupervised methods for comparison, including PCA-KM (Celik 2009), DSFA (Du et al. 2019), DCVA (Saha, Bovolo, and Bruzzone 2019), GMCD (Tang et al. 2021) and CVA (Bovolo and Bruzzone 2006). In addition, AnyChange (Zheng et al. 2024) is used for comparison, considering single-class CD as binary CD. The closest to our task is SCM (Tan et al. 2023), which is designed for unsupervised building CD.

Train on: \ Test on:	LEVIR-CD		WHU-CD		S2Looking	
	IoU^c	F_1^c	IoU^c	F_1^c	IoU^c	F_1^c
LEVIR-CD	84.9	91.8	50.3	66.9	1.5	2.9
WHU-CD	16.9	28.9	87.9	93.5	3.1	6.0
S2Looking	45.6	62.7	22.0	36.1	49.6	66.3
OVCD (ours)	53.5	69.7	61.1	75.8	23.9	38.5
Δ	↑**7.9**	↑**7.0**	↑**10.8**	↑**8.9**	↑**11.9**	↑**17.1**

Table 3: Comparison of OVCD method and supervised CD method on cross-datasets (unseen data). Gray indicates training and evaluation on the same dataset, which can be seen as the upper bound and is not involved in the comparison. "ours" indicates the best results reported in Table 1.

Results and Discussion

Building CD. As listed in Table 1, we build several OVCD methods under the M-C-I and I-M-C frameworks. For the M-C-I framework, we use SAM or SAM 2 as the mask proposal, DINO or DINOv2 as the comparator, and SegEarth-OV as the identifier. Their performance differences are not

significant, which suggests that the M-C-I framework is relatively stable, and this stability stems from the strong generalization capabilities of SAMs, DINOs and CLIPs (on which SegEarth-OV relies). The performance of the I-M-C methods is highly dependent on the identifier. According to their reports (Zhao et al. 2024; Shen et al. 2024), both MM-Grounding-DINO and APE are trained only on some public general image data, but the latter has significantly superior ability in remote sensing scenarios than the former, leading to the performance difference in the OVCD task. Compared to the S2Looking dataset, LEVIR-CD and WHU-CD are simpler and they have higher image quality and clearer context. An interesting observation is that in simple data, the methods under the I-M-C framework achieve the best performance, while in complex data, the M-C-I methods achieve the best results. This suggests that the M-C-I framework is more robust to some extent, and in some complex scenarios, suboptimal identifiers in the I-M-C framework result in serious error accumulation.

Land-cover CD. Beyond buildings, there are several other land cover types that are of interest to users. In Table 2, we evaluate the OVCD of 6 categories in the SECOND dataset. It can be observed that the M-C-I methods generally outperform the I-M-C methods due to the fact that the identifiers in I-M-C can hardly recognize some categories that are often regarded as *"background"* in natural images, *e.g.*, *"Tree"*. In addition, some categories, *e.g.*, *"Low vegetation"*, *"N.v.g. surface"*, are difficult to instantiate. All methods yield superior results for *"Building"* compared to other categories. We suppose that this is because buildings, as one of the most common man-made land covers, have stronger *"foreground"* characteristics and a certain data bias in the training data of the foundation models.

Effectiveness of OVCD. The significance of OVCD is to bridge the gap between vision and language, improve the generalization of the model, and enable the model to have the ability to detect any change of interest, thus avoiding costly re-training. Therefore, in Table 3, we compare the cross-dataset performance of supervised learning models with the OVCD method (the best results reported in Table 1). We select the state-of-the-art supervised CD model, Changer (Fang, Li, and Li 2023; Li et al. 2024c). It can be observed that the models trained on LEVIR-CD and WHU-CD are nearly unusable on S2Looking. The models trained on S2Looking perform reasonably well on other datasets, but are still far from the results trained on their own datasets. The OVCD method is significantly superior to the best cross-dataset results on all datasets, which confirms the availability and potential of OVCD in real-world scenarios.

Visualization. In Figure 3, we visualize some OVCD results for the LEVIR-CD, WHU-CD and SECOND datasets. It can be observed that M-C-I method can handle small targets well without pre-cropping, but there are some compact targets that still cannot be instantiated and detected. The I-M-C method is highly dependent on the capability of the identifier, which ensures high precision, but some targets may be dropped at the initial stage (low recall).

Change Detection for Novel Classes. In Figure 4, we explore the novel class detection capability of the proposed method. "Photovoltaic panel" is an uncommon category in CD task. OVCD achieves its CD without the need for data collection and annotation or model retraining. In addition, we also attempt to detect changes of special targets, such as "blue houses". The results on the right side of Figure 4 show the precision of the proposed method, without false positives for other types of houses.

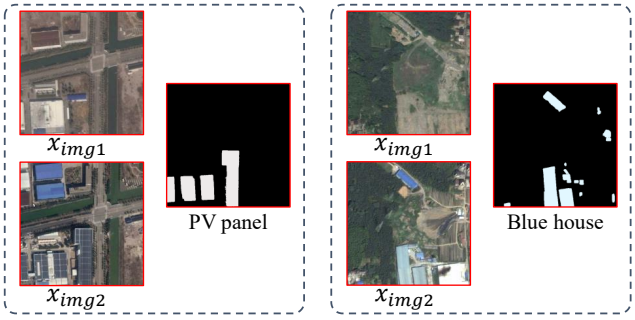

Figure 4: Detection of uncommon changes using the proposed OVCD method.

Design of Prompts. In inference, VLMs are sensitive to textual prompt inputs. Consider a binary segmentation scene where we generally use the template with the category name as the foreground prompt and the one with *"background"* as the background prompt. Indeed, this works for the common category of CD, *i.e. "Building"*. However, for the various categories in the land cover, such simple prompts may not work well. We can alleviate this issue by adding synonyms to the foreground prompt and other categories involved in the dataset to the background prompt. However, carefully designing prompt words for each category is not what we expect. Possible solutions to improve this issue are to design a comprehensive vocabulary dictionary or to use LLM to design prompts (Wysoczańska et al. 2024).

Conclusion

In this paper, we propose a new task, open-vocabulary change detection, which realizes the connection between vision and language in CD and can segment any change of interest. Inspired by the *"4W"* assertion, we propose two frameworks for OVCD, M-C-I and I-M-C. Based on these two frameworks, we instantiate several training-free OVCD methods equipped with off-the-shelf VLMs. Through comprehensive experiments on multiple CD datasets, we show the superiority of the proposed OVCD method over previous unsupervised methods and further demonstrate the effectiveness of OVCD in practice. In addition, we contribute the first OVCD codebase, DynamicEarth, to the Earth vision community for algorithm development, evaluation, and application. Although the proposed OVCD method still falls short of purely supervised methods, we believe that open-world perception is what is needed for practical CD. We hope that subsequent research will further improve the OVCD method, either training-based or training-free, either in accuracy or efficiency, *etc*.

Acknowledgements

This work is supported by the National Key R&D Program of China (2022YFA1004100), National Natural Science Foundation of China (No. 62272375, No. 62425113, No. 62192781), Major Key Project of PCL (PCL2024A06), and Tianyuan Fund for Mathematics of the National Natural Science Foundation of China (No. 12426105).

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Bovolo, F.; and Bruzzone, L. 2006. 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.

Cao, Q.; Chen, Y.; Ma, C.; and Yang, X. 2024. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*.

Celik, T. 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE geoscience and remote sensing letters*, 6(4): 772–776.

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

Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; and Shi, Z. 2024. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cho, S.; Shin, H.; Hong, S.; Arnab, A.; Seo, P. H.; and Kim, S. 2024. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4123.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Ding, L.; Zuo, X.; Hong, D.; Guo, H.; Lu, J.; Gong, Z.; and Bruzzone, L. 2025. S2C: Learning Noise-Resistant Differences for Unsupervised Change Detection in Multimodal Remote Sensing Images. *arXiv preprint arXiv:2502.12604*.

Du, B.; Ru, L.; Wu, C.; and Zhang, L. 2019. 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.

Fang, S.; Li, K.; and Li, Z. 2023. Changer: Feature interaction is what you need for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.

Howarth, P. J.; and Wickware, G. M. 1981. Procedures for change detection using Landsat digital data. *International Journal of Remote Sensing*, 2(3): 277–291.

Ji, S.; Wei, S.; and Lu, M. 2018. 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.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2024. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, K.; Cao, X.; Deng, Y.; Song, J.; Liu, J.; Meng, D.; and Wang, Z. 2024a. SemiCD-VL: Visual-Language Model Guidance Makes Better Semi-supervised Change Detector. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, K.; Cao, X.; and Meng, D. 2024. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Li, K.; Dong, F.; Wang, D.; Li, S.; Wang, Q.; Gao, X.; and Chua, T.-S. 2024b. Show me what and where has changed? question answering and grounding for remote sensing change detection. *arXiv preprint arXiv:2410.23828*.

Li, K.; Jiang, J.; Codegoni, A.; Han, C.; Deng, Y.; Chen, K.; Zheng, Z.; Chen, H.; Zou, Z.; Shi, Z.; et al. 2024c. OpenCD: A Comprehensive Toolbox for Change Detection. *arXiv preprint arXiv:2407.15317*.

Li, K.; Liu, R.; Cao, X.; Bai, X.; Zhou, F.; Meng, D.; and Wang, Z. 2025a. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10545–10556.

Li, K.; Wang, D.; Wang, T.; Dong, F.; Zhang, Y.; Zhang, L.; Wang, X.; Li, S.; and Wang, Q. 2025b. RSVG-ZeroOV: Exploring a Training-Free Framework for Zero-Shot Open-Vocabulary Visual Grounding in Remote Sensing Images. *arXiv preprint arXiv:2509.18711*.

Li, K.; Wang, D.; Xu, H.; Zhong, H.; and Wang, C. 2024d. Language-guided progressive attention for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2025. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Lu, D.; Mausel, P.; Brondizio, E.; and Moran, E. 2004. Change detection techniques. *International journal of remote sensing*, 25(12): 2365–2401.

Mahmoudzadeh, H. 2007. Digital change detection using remotely sensed data for monitoring green space destruction in Tabriz.

Malila, W. A. 1980. Change vector analysis: An approach for detecting forest changes with Landsat. In *LARS symposia*, 385.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peng, D.; Liu, X.; Zhang, Y.; Guan, H.; Li, Y.; and Bruzzone, L. 2025. Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges. *International Journal of Applied Earth Observation and Geoinformation*, 136: 104282.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren; and Malik. 2003. Learning a classification model for segmentation. In *Proceedings ninth IEEE international conference on computer vision*, 10–17. IEEE.

Ren, S.; Luzi, F.; Lahrichi, S.; Kassaw, K.; Collins, L. M.; Bradbury, K.; and Malof, J. M. 2024a. Segment anything, from space? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8355–8365.

Ren, T.; Chen, Y.; Jiang, Q.; Zeng, Z.; Xiong, Y.; Liu, W.; Ma, Z.; Shen, J.; Gao, Y.; Jiang, X.; et al. 2024b. DINO-X: A Unified Vision Model for Open-World Object Detection and Understanding. *arXiv preprint arXiv:2411.14347*.

Saha, S.; Bovolo, F.; and Bruzzone, L. 2019. Unsupervised deep change vector analysis for multiple-change detection in VHR images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6): 3677–3693.

Shen, L.; Lu, Y.; Chen, H.; Wei, H.; Xie, D.; Yue, J.; Chen, R.; Lv, S.; and Jiang, B. 2021. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24): 5094.

Shen, Y.; Fu, C.; Chen, P.; Zhang, M.; Li, K.; Sun, X.; Wu, Y.; Lin, S.; and Ji, R. 2024. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13193–13203.

Shi, J.; and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8): 888–905.

Tan, X.; Chen, G.; Wang, T.; Wang, J.; and Zhang, X. 2023. Segment Change Model (SCM) for Unsupervised Change detection in VHR Remote Sensing Images: a Case Study of Buildings. *arXiv preprint arXiv:2312.16410*.

Tang, X.; Zhang, H.; Mou, L.; Liu, F.; Zhang, X.; Zhu, X. X.; and Jiao, L. 2021. 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.

Tian, S.; Ma, A.; Zheng, Z.; and Zhong, Y. 2020. Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery. *arXiv preprint arXiv:2011.03247*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wysoczańska, M.; Vobecky, A.; Cardiel, A.; Trzciński, T.; Marlet, R.; Bursuc, A.; and Siméoni, O. 2024. A Study of Test-time Contrastive Concepts for Open-world, Open-vocabulary Semantic Segmentation. *arXiv preprint arXiv:2407.05061*.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4818–4829.

Yang, K.; Xia, G.-S.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; and Zhang, L. 2021. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Ye, C.; Zhuge, Y.; and Zhang, P. 2025. Towards Open-Vocabulary Remote Sensing Image Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhao, X.; Chen, Y.; Xu, S.; Li, X.; Wang, X.; Li, Y.; and Huang, H. 2024. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*.

Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; and Wang, J. 2023. Fast segment anything. *arXiv preprint arXiv:2306.12156*.

Zheng, Z.; Zhong, Y.; Zhang, L.; and Ermon, S. 2024. Segment Any Change. In *Advances in Neural Information Processing Systems*.

Zhu, C.; and Chen, L. 2024. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.