S2C: A Noise-Resistant Difference Learning Framework for Unsupervised Change Detection in VHR Remote Sensing Images

Lei Ding, Xibing Zuo, Haitao Guo, Jun Lu*, Zhihui Gong, Xuanguang Liu, Jicang Lu

Information Engineering University, Zhengzhou, China.

Abstract

Unsupervised Change Detection (UCD) in Very High Resolution (VHR) Remote Sensing (RS) images remains to be a difficult challenge due to the inherent spatio-temporal complexity within data. Inspired by recent advancements in Visual Foundation Models (VFMs) and Contrastive Learning (CL), this research aims to develop CL methodologies to translate implicit knowledge in VFM into change representations, thus eliminating the need for explicit supervision. To this end, we introduce a Semantic-to-Change (S2C) learning framework for UCD in VHR RS images. Differently from existing CL methodologies that typically focus on learning multi-temporal similarities, we introduce a novel triplet learning strategy that explicitly models temporal differences, which are crucial to the CD task. Furthermore, random spatial and spectral perturbations are introduced during training to enhance robustness to temporal noise. In addition, a grid sparsity regularization is defined to suppress insignificant changes, and an IoU-matching algorithm is developed to refine the CD results. Experiments on three benchmark CD datasets demonstrate that the proposed S2C learning framework achieves significant improvements in accuracy, surpassing current state-of-the-art by over 31%, 9% and 23%, respectively. It also demonstrates robustness and sample efficiency, suitable for training and adaptation of various VFMs or backbone neural networks.

Code — https://github.com/DingLei14/S2C

Extended version — https://arxiv.org/abs/2502.12604

Introduction

Change detection (CD) in Remote Sensing (RS) is the process of identifying and segmenting regions of change using multi-temporal observations of the same geographic area. Over the past decade, great advances have been achieved in CD utilizing deep learning (DL) techniques. State-Of-The-Art (SOTA) methodologies (Chen, Qi, and Shi 2021; Ding et al. 2024) have obtained accuracy levels that exceed 90% in the F_1 metric in various benchmark datasets for CD. However, most majority of these DL-based CD methods require large amounts of high-quality labeled data, which are difficult to collect due to the scarcity of change samples. Consequently, the deployment of CD algorithms in real-world applications still faces significant challenges.

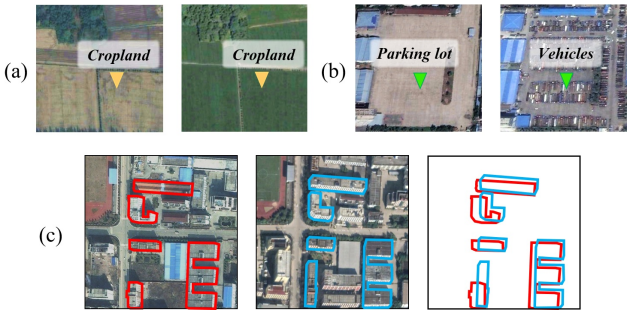

Figure 1: The major types of temporal noise in CD of VHR RSIs include: (a) spectral variations, (b) insignificant changes and (c) spatial misalignment.

To reduce dependence on training data, an increasing number of studies have been conducted on unsupervised change detection (UCD) in recent years(Ding et al. 2025). However, most of these investigations focus on UCD of medium-resolution RS images (Chen and Bruzzone 2021). UCD of high-resolution RS images presents greater challenges to CD algorithms due to the increased spatial and temporal complexity. Fig.1 illustrates several instances of noise encountered in the CD of HR RS images, including: i) spectral variations. This can be attributed to either seasonal variations in vegetation or disparities in imaging sensors and illumination conditions. ii) Insignificant changes. In CD applications, only specific changes are of interest, such as building changes in urban management and cropland changes in agriculture monitoring. Certain temporary changes, such as the parked vehicles depicted in Fig.1(b), are often deemed irrelevant noise. iii) Spatial misalignment. This may arise from varying imaging angles, optical distortion, or errors in image registration. Consequently, deep neural networks (DNNs) encounter significant challenges in learning to differentiate between semantic changes and temporal differences in an unsupervised manner. In recent studies, contrastive learning (CL)(Sohn et al. 2020) and Visual Foundation Models (VFMs) (Kirillov et al. 2023) are identified as two effective techniques to mitigate data dependence in CD. The former leverages the intrinsic con-

*Corresponding author.

sistency within data, while the latter incorporates external knowledge to learn generalizable semantic representations. Advances have been achieved in semi-supervised CD of HR RSIs leveraging CL (Yang et al. 2023; Chen and Bruzzone 2021) and VFMs (Ding et al. 2024; Zheng et al. 2024). However, UCD of HR RSIs using either CL or VFM remain challenging due to the inherent spatio-temporal complexity within this task. In CL-based CD, various studies follow a consistency regularization framework(Sohn et al. 2020). This approach significantly enhances the generalization and robustness of feature representations, yet still requires a certain proportion of training data. VFM-based CD often employ VFMs, such as the Segment Anything Model (SAM) (Kirillov et al. 2023), to exploit semantic features and decode change masks (Zheng et al. 2024). Nonetheless, a significant domain gap exist between the training domains of VFMs and RS images (Ji et al. 2024), adversely affecting their recognition capabilities. Furthermore, accurate translation of semantic features into change maps still requires certain degree of supervision.

In this article, we explore the integration of CL and VFM to accomplish UCD in VHR RS images. It is observed that these two methodologies effectively complement each other: CL provides self-supervised training objectives essential for adapting VFMs to the RS domain and for mapping the changes. In contrast, VFMs embed pixel-level semantic representations, a capability that is usually absent in typical CL frameworks. Furthermore, we extend the existing CL frameworks to incorporate the spatio-temporal correlations unique to CD. The resulting methodologies have led to substantial accuracy improvements over state-of-the-art (SOTA) methods. The major technical contributions in this study can be summarized as follows:

1) Developing a UCD framework that explicitly models unsupervised semantic changes. To the best of our knowledge, the proposed S2C framework is the first to incorporate VFM into CL for CD. It integrates multiple innovative designs, including CL, VFM, Low-Rank Adaptation (LoRA), and an IoU refinement algorithm.

2) Introducing two novel CL paradigms for CD: Consistency-regularized Temporal Contrast (CTC) and Consistency-regularized Spatial Contrast (CSC), to capture multi-temporal differences and consistency. CTC introduces a triplet learning strategy to explicitly model semantic changes, complementing prior work focused on similarities.

3) Proposing a grid sparsity regularization to promote sparse and compact change mapping. It is executed at grid scales to avoid training collapsion and ensure efficiency.

Related Work

Unsupervised Change Detection

Unsupervised CD poses great challenges to DL-based approaches due to the absence of explicit supervision(Liu et al. 2025). To address this challenge, three principal strategies have been developed, including feature difference mapping, generative representation, and knowledge transfer.

Difference mapping is an essential step in CD that transforms deep features into change representations. Classic methods for difference mapping include principal component analysis (PCA) (Bruzzone and Prieto 2000; Gao et al. 2016), change vector analysis (CVA) (Saha, Bovolo, and Bruzzone 2019; Wu et al. 2022), slow feature analysis (Wu, Du, and Zhang 2014; Du et al. 2019), half-sibling regression (Kondmann et al. 2022), etc. Due to absence of explicit supervision, some methods (Saha, Bovolo, and Bruzzone 2019) directly employ features extracted by deep neural networks (DNNs) (without fine-tuning), resulting in weak semantic abstraction. Meanwhile, several works employ image generation methods to reduce style differences between multi-temporal observations, a strategy known as generative transcoding. It enables CD between heterogeneous inputs such as optical and synthetic aperture-radar (SAR) images (Saha, Bovolo, and Bruzzone 2020). The frequently used generative approaches include auto-encoder (Chen et al. 2022) and Generative Adversarial Networks (GANs) (Noh et al. 2022). In (Wu, Du, and Zhang 2023) a generative framework is introduced to iteratively optimize CD results.

Contrastive Learning

Contrastive Learning (CL), a type of self-supervision approaches, constructs and compares positive and negative pairs to exploit the semantic consistency in unlabeled data. An established paradigm of CL in visual recognition is to introduce weak-to-strong perturbations, thus regularizing DNNs to learn robust semantic representations (Sohn et al. 2020). These perturbations can be introduced into input images or embedded features (Yang et al. 2023).

In CD, bi-temporal images of the same/different regions are often utilized to construct contrastive pairs. In (Chen and Bruzzone 2021) a simple CL paradigm for CD is introduced, where change pairs are constructed with cropped RSIs at same vs. different locations. Differently, Bandara et al. (Bandara and Patel 2022) introduce perturbations on the bi-temporal difference features and perform consistency-regularized CL. In (Mall, Hariharan, and Bala 2023) CL for CD in long-term temporal observations is introduced.

Overall, CL is mostly utilized in semi-supervised CD (Yang et al. 2023; Bandara and Patel 2022) and weakly-supervised CD (Zhao et al. 2024) to leverage the sparse supervision signals available. Literature methods on CL-based UCD predominately exploit the temporal similarity embeddings (Chen and Bruzzone 2021), yet there exists a notable gap in exploration of temporal difference embeddings. In this study, we investigate the joint exploitation of temporal consistency and difference embeddings with CL.

VFM-Based Change Detection

In recent years, there is a trend towards developing Visual Foundation Models (VFMs), such as CLIP (Radford et al. 2021) and Segment Anything Model (SAM) (Kirillov et al. 2023), to acquire comprehensive recognition capabilities. VFMs are trained on web-scale image datasets to capture universal features applicable to various tasks. However, since the VFMs are mostly trained in common natural scenes, they demonstrate bias when applied to recognition of RS images (Ji et al. 2024). Considering the spectral and temporal characteristics of RSIs, several RS foundation models

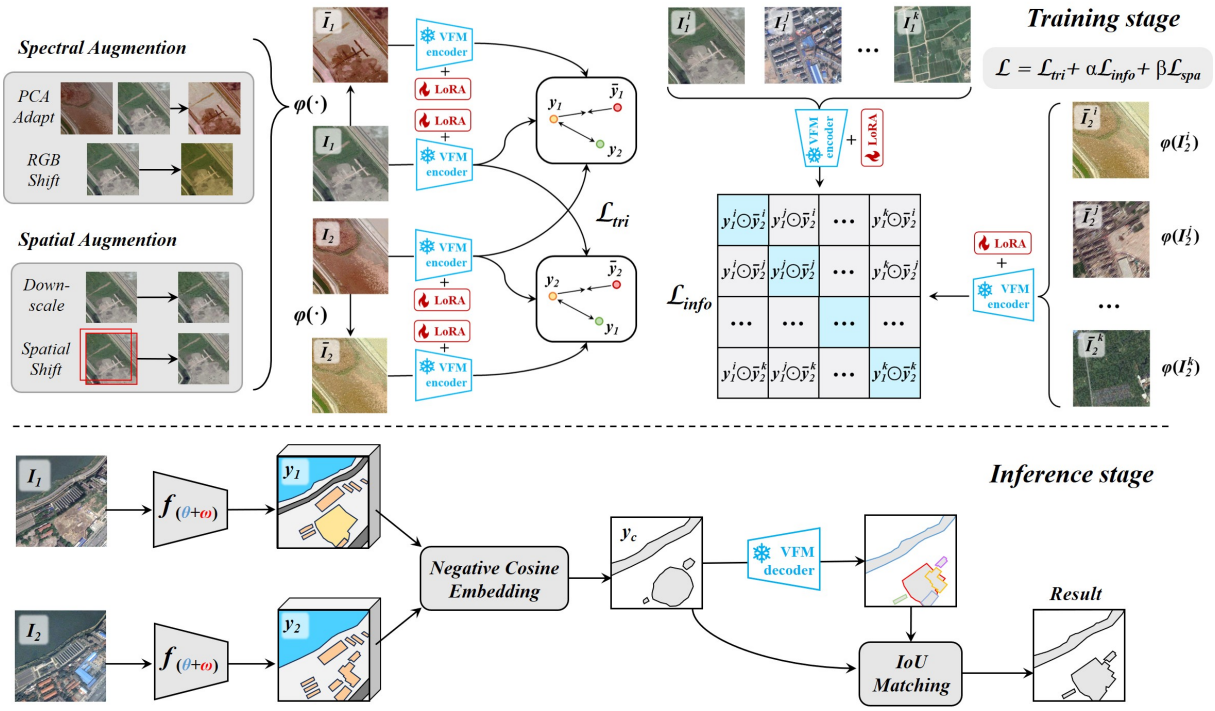

Figure 2: Overview of the proposed S2C framework for UCD. Triplet losses are calculated with bitemporal images and their augmented copies to learn temporal differences; discriminative losses are calculated between bitemporal images of different regions to learn temporal consistency. Random perturbations are introduced to simulate the spectral and spatial variations.

(FMs) have been developed, including SpectralGPT (Hong et al. 2024) and SkySense (Guo et al. 2024). However, employing these models for CD still necessitates incorporating and fine-tuning CD-specific modules.

Recent paradigms explore employing VFMs to achieve sample-efficient CD. SAM-CD(Ding et al. 2024) first adapt VFMs to CD of RS images, obtains superior accuracy over fully-supervised CD methods and demonstrates label efficiency in the training. In (Wang, Zhang, and Shi 2023), SAM is utilized to generate pseudo labels, using vague change maps as prompts. Chen at el. (Chen, Song, and Yokoya 2024) employ SAM to achieve unsupervised CD between optical images and map data. In (Zheng et al. 2024), zero-shot CD is achieved by measuring the similarity of SAM-encoded features. In (Dong et al. 2024) CLIP is employed to learn visual language representations to improve CD accuracy. Recently, Li et al. propose an I-M-C framework utilizing VFMs to achieve Open-Vocabulary CD (Li et al. 2025).

Despite these previous works leveraging VFM for CD, UCD on VHR RSIs is still challenging and the SOTA accuracy is limited. In this research, we explore VFM-based UCD by incorporating CL to replace explicit supervision.

Proposed Approach

Overview of the S2C framework

DL-based CD essentially learns to project multi-temporal RS images I_1, I_2 into a binary change map y_c. Let f_θ denote an encoding function parameterized by θ, and g being a projection function, this process can be formulated as:

$$\mathbf{y_1} = f_\theta(I_1), \mathbf{y_2} = f_\theta(I_2), \mathbf{y_c} = g(\mathbf{y_1}, \mathbf{y_2}) \quad (1)$$

where $\mathbf{y_1}, \mathbf{y_2} \in \mathbb{R}^{s\times h\times w}$ are embedded semantic latent.

The proposed S2C architecture, depicted in Fig.2, is a UCD framework that initially conducts self-supervised CL to exploit task-specific semantic features. Subsequently, these semantic representations are translated into CD results through projecting and refinement algorithms. Although this framework can train DNNs from scratch, using VFM as feature encoders imperatively leads to better accuracy. Therefore, we fine-tune a VFM with additional parameters w introduced to adapt it to the RS domain. We adopt the LoRA (Hu et al. 2022) (with $rank = 4$) to learn to adapt the VFM parameters, which correspond to the encoding function of $f_{\theta+w}$. The VFM can be any off-the-shelf models, as its inner structure is not modified in our S2C framework (see ablation on different VFMs in Sec.). In the training phase, θ is frozen to retain the pre-trained visual knowledge, while w are the LoRA weights trained with CL paradigms to exploit temporal semantic features.

The training process is conducted within two CL paradigms to learn the semantic representations relevant to CD. Two CL paradigms are introduced to embed difference and consistency representations, respectively, which are elaborated in *Experiments*. The associated loss functions are $\mathcal{L}_{tri}$ and $\mathcal{L}_{info}$, respectively. In addition, we further introduce a compactness regularization objective to learn sparse and compact change representations, noted as $\mathcal{L}_{spa}$.

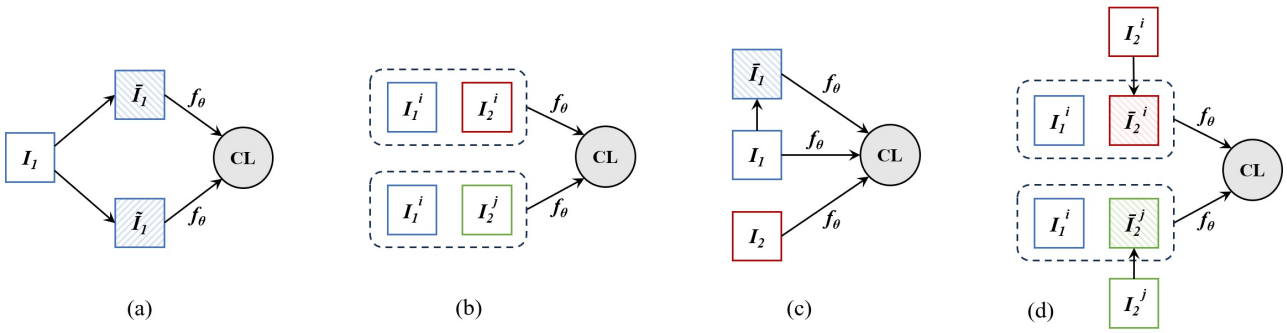

Figure 3: Comparison on CL paradigms in CD. (a) **Consistency regularization**: f_θ extracts stable representations across weak/strong perturbations; (b) **Spatial contrast**: f_θ distinguishes same/different regions; (c) Proposed **Consistency-regularized Temporal Contrast** (CTC): f_θ identifies temporal differences independent of appearance variations, and (d) Proposed **Consistency-regularized Spatial Contrast** (CSC): f_θ distinguishes same/different regions despite perturbations.

The joint training objective is:

$$\mathcal{L} = \mathcal{L}_{tri} + \alpha\mathcal{L}_{info} + \beta\mathcal{L}_{spa} \tag{2}$$

where α and β are two weighting parameters.

In the inference phase, the semantic latent y_1 and y_2 are first mapped into a coarse change map, then refined using the VFM decoder and an IoU matching function. The details are elaborated in *Supplementary Materials*.

Contrastive Change Learning

Before introducing the proposed CL paradigms, let us first review the two typical CL paradigms in CD, and analyze their usage and limitations.

1) Consistency regularization (CR). As depicted in Fig.3(a), a DNN f_θ learns more robust and generalizable feature representations. An image I is first augmented with weak and strong transformations, resulting in two copies $\tilde{I}$ and $\bar{I}$. Then a distance loss is calculated between the two copies to ensure consistency across perturbations.

Since this learning paradigm does not explicitly model differences/similarities, it is often adopted in semi-supervised (Bandara and Patel 2022) or weakly supervised (Zhao et al. 2024) learning settings to extend the CD insights learned with limited samples.

2) Spatial contrast (SC). As illustrated in Fig.3(b), f_θ learns to differentiate between bitemporal image pairs $[I_1^i, I_2^i]$ of the same region i and $[I_1^i, I_2^j]$ of different regions i and j. This drives f_θ to learn consistent embeddings against temporal variations. Areas with high similarity are identified as $unchange$, whereas their opposites are detected as $change$ (Chen and Bruzzone 2021).

However, we identify several limitations in this paradigm: i) Changes are identified through negative embedding of similarities rather than through explicit modeling. This often causes sensitivity to noise. ii) f_θ focus on discriminative elements within a region, such as certain edges or corners, rather than effectively exploiting the local semantic context.

Considering these limitations, we subsequently introduce two innovative CL paradigms specifically tailored to the context of CD.

Consistency-regularized Temporal Contrast (CTC). An RS image I_1 is first augmented with a transform function $\phi(\cdot)$, producing a copy $\bar{I}_1$. Subsequently, I_1 is employed as an anchor for comparison with both a positive sample $\bar{I}_1$ and a negative sample I_2. $\phi(\cdot)$ simulates spectral and spatial noise between multi-temporal observations, as illustrated in Fig.1. Consequently, f_θ learns to exploit noise-invariant difference representations, i.e., semantic changes.

With greater details, Fig.2 illustrates the CTC paradigm with bi-directional comparisons within $[I_1, \bar{I}_1, I_2]$ and $[I_2, \bar{I}_2, I_1]$. The transformation function $\phi(\cdot)$ comprises a series of spatial and spectral augmentations executed randomly at each training iteration, including *random shifting*, *down-sampling*, *RGBshift*, and a *PCA adaptation* which adapts the spectral distribution of the positive sample to approach that of the negative sample. The spatial operations are performed to simulate spatial misalignment and imaging degradation/distortion, while the spectral operations replicate imaging and seasonal variations. Collectively, these transformations enhance the algorithm's robustness against temporal noise.

A triplet training objective $\mathcal{L}_{tri}$ using cosine distance is utilized for comparisons within the triplets. This is to align with the cosine difference embedding during the inference stage. The calculations are as follows:

$$\begin{aligned}\mathcal{L}_{tri} = &max[cos(\mathbf{y_1}, \mathbf{y_2}) - cos(\mathbf{y_1}, \bar{\mathbf{y}}_\mathbf{1}) + m, 0] \\ &+max[cos(\mathbf{y_2}, \mathbf{y_1}) - cos(\mathbf{y_2}, \bar{\mathbf{y}}_\mathbf{2}) + m, 0]\end{aligned} \tag{3}$$

where $m = 1$ is a margin parameter to promote separation between the anchor and positive.

Consistency-regularized Spatial Contrast (CSC). This contrastive learning paradigm integrates CR into typical SC learning, thereby enhancing the embedding of spatial consistency against perturbations. CSC alleviates the vulnerability to noise inherent in the SC paradigm by incorporating transformation $\phi(\cdot)$. The transformations, particularly the spatial transformations, reduce dependence on high-frequency spatial details, thereby necessitating the exploitation of local semantic contexts (such as color and texture patterns).

We have introduced an additional variation in CSC, i.e., the calculation of consistency at each spatial posi-

tion. Given a batch consisting of N paired RS images $\{[I_1^i, I_2^i], [I_1^j, I_2^j], ..., [I_1^k, I_2^k]\}$, we first apply $\phi(\cdot)$ on each of the temporal images, thus getting two sets of augmented images. These images are further encoded with $f_{\theta+w}$, resulting in 4 sets of features: $[\mathbf{y_1^i}, \mathbf{y_1^j}, ..., \mathbf{y_1^k}]$, $[\mathbf{y_2^i}, \mathbf{y_2^j}, ..., \mathbf{y_2^k}]$, $[\bar{\mathbf{y}}_\mathbf{1}^\mathbf{i}, \bar{\mathbf{y}}_\mathbf{1}^\mathbf{j}, ..., \bar{\mathbf{y}}_\mathbf{1}^\mathbf{k}]$ and $[\bar{\mathbf{y}}_\mathbf{2}^\mathbf{i}, \bar{\mathbf{y}}_\mathbf{2}^\mathbf{j}, ..., \bar{\mathbf{y}}_\mathbf{2}^\mathbf{k}]$. We then calculate the co-occurrences between them, resulting in two matrices each with $N \times N$ dimensions, as illustrated in Fig.3. We utilize an infoNCE loss function to effectively train $f_{\theta+w}$ to differentiate genuine image pairs. It is calculated across both temporal phases, represented as:

$$\begin{aligned}\mathcal{L}_{info} = &-\frac{1}{N}\sum_{u=1}^{N}\log\left[\frac{\exp\left(\mathbf{y}_1^u \odot \bar{\mathbf{y}}_2^u\right)}{\sum_{v=1}^{N}\exp\left(\mathbf{y}_1^u \odot \bar{\mathbf{y}}_2^v\right)}\right] \\ &-\frac{1}{N}\sum_{u=1}^{N}\log\left[\frac{\exp\left(\mathbf{y}_2^u \odot \bar{\mathbf{y}}_1^u\right)}{\sum_{v=1}^{N}\exp\left(\mathbf{y}_2^u \odot \bar{\mathbf{y}}_1^v\right)}\right]\end{aligned} \tag{4}$$

where $\odot$ denotes a novel similarity measurement function that we introduce in this study. Instead of pooling the spatial features into single vectors for similarity calculation (Chen and Bruzzone 2021), we compute the similarity at each spatial position p, denoted as:

$$\mathbf{y} \odot \bar{\mathbf{y}} = \frac{1}{w \times h}\sum_{p}\left(\frac{\mathbf{y}^p \cdot \bar{\mathbf{y}}^p}{|\mathbf{y}^p||\bar{\mathbf{y}}^p|}\right) \tag{5}$$

Both $\mathcal{L}_{tri}$ and $\mathcal{L}_{info}$ are calculated based on cosine similarity. While $\mathcal{L}_{tri}$ embeds appearance-invariant temporal differences, $\mathcal{L}_{info}$ embeds noise-resilient temporal consistencies. Therefore, when certain temporal consistency patterns are captured in CSC, it suppresses the difference representations of a same area in CTC.

Grid Sparsity loss

Changed items are commonly sparsely distributed in RS images, each manifested as a compact segment. In contrast, edges and points are often noise. Although training objectives that promote sparse representations have been explored in the literature, they typically calculate and penalize the average value of $\mathbf{y_c}$ (Bandara and Patel 2023). However, this approach does not guarantee sparsity, as there exists a trivial solution to learn an additional bias term on $\mathbf{y_c}$.

Differently, we propose a novel grid sparsity loss where sparsity is assessed at the level of each local grid rather than at each pixel. Considering the frequency of changes along with the spatial resolution in an RS image, we first define a sparsity threshold T and a grid size d. Subsequently, the average density of each grid $\mathbf{g}$ is calculated and ranked, while a $1-T$ ratio of grids with the lowest density is selected for loss calculation.

$$\begin{aligned}y^{\mathbf{g}} &= \frac{1}{d*d}\sum_{p\in\mathbf{g}}\mathbf{y}_c^p, n = wh*(1-T)/d^2, \\ \mathcal{L}_{spa} &= max\{\frac{1}{n}\sum^{n}[sort\uparrow(y^{\mathbf{g}})], 0\}\end{aligned} \tag{6}$$

where we empirically set $d = 16$ for VHR RS images, and using $T = 0.2$ for data with sparse changes. This objective ensures that less than a $1-T$ proportion of potential changes exhibit high values, whereas the insignificant change representations in other areas are minimized.

Change mapping

In the training phase, VFM and CL are employed to enhance exploitation of semantic contexts across multi-temporal image domains. During inference, the major challenge lies in accurately mapping fine-grained changes. We employ a coarse-to-fine refinement strategy. First, a coarse change probability map $\mathbf{y}_c$ is derived by projecting the negative cosine embedding of the bi-temporal semantic embeddings:

$$\mathbf{y}_c = \sigma[-cos(\mathbf{y}_1, \mathbf{y}_2) * \eta] \tag{7}$$

where σ is a $sigmoid$ function, $\eta = ln(1/0.07)$ is a scaling factor defined following literature practice.

Then, we employ a pretrained VFM decoder g_γ to segment two groups of bi-temporal masks $M_1 = \{m_1^1, m_1^2, ..., m_1^k\}$ and $\{M_2 = m_2^1, m_2^2, ..., m_2^k\}$ using the spatial prompts generated on high-response regions in $\mathbf{y}_c$. Given the logic implication of *change*, high-overlap objects in M_1 and M_2 can be inferred as false alarms. Therefore, we conduct an XOR-alike matching algorithm (denoted $\oplus$) to merge M_1, M_2 and eliminate the overlaps:

$$M_{12} = M_1 \oplus M_2 \tag{8}$$

We further implement an Intersection-over-Union (IoU) analysis between $\mathbf{y}_c$ and M_{12} to match the VFM-generated masks with the high-confidence regions in $\mathbf{y}_c$. The matched objects replace their counterparts in $\mathbf{y}_c$ as the changed items. This refines the coarse change masks with the VFM decoded change instances with refined spatial details. For more details, readers are encouraged to find the relevant implementations that will be made available online.

Experiments

Datasets and Evaluation Metrics

Datasets	Platform	Resolution	Image size	Data Size	Change Type
CLCD	satellite	0.5-2m	512×512	600	agricultural
SECOND	aerial	0.5-3m	512×512	4,662	land cover
Levir	satellite	0.5m	1024×1024	637	building

Table 1: Statistical summary of the experimental datasets.

To test the efficacy of the proposed methodologies, experiments are conducted on three benchmark datasets with varied data distributions and semantic annotations, i.e., CLCD (Liu et al. 2022), SECOND (binary) (Yang et al. 2022) and Levir (Chen and Shi 2020). Table 1 presents an overview of each dataset. Since the experimental methods are unsupervised, we do not use any label within the train set. Notably, the Levir dataset focuses solely on building changes, while CLCD and SECOND encompass various type of changes. Its larger spatial size and sparser change instances make it more challenging in the context of UCD.

We adopt the most commonly used metrics in binary CD, including Overall accuracy (*OA*), Precision (*Pre*), Recall

Methods	Backbone	Accuracy (%)			
		OA	Pre	Rec	F_1
effi.SAM + CVA	effi.SAM (vit-t)	61.24	13.90	81.01	23.73
effi.SAM + SC	effi.SAM (vit-t)	92.23	45.76	24.07	31.55
S2C (CSC only)	effi.SAM (vit-t)	90.93	36.55	29.81	32.84
S2C (CTC only)	effi.SAM (vit-t)	85.98	28.25	57.40	37.86
S2C (CSC + $\mathcal{L}_{spa}$)	effi.SAM (vit-t)	89.71	33.35	38.33	35.67
S2C (CTC + $\mathcal{L}_{spa}$)	effi.SAM (vit-t)	91.06	39.52	38.03	38.76
S2C (CSC + CTC)	effi.SAM (vit-t)	87.35	31.01	57.12	40.19
S2C	effi.SAM (vit-t)	90.57	39.04	47.65	42.92
S2C	effi.SAM (vit-s)	89.58	37.51	**60.21**	46.22
S2C + *IoU refine*	effi.SAM (vit-s)	**91.47**	**43.85**	52.28	**47.69**
S2C (w/o. VFM)	ResNet18	87.75	26.80	37.34	31.21
S2C (w/o. VFM)	ResNet34	86.28	28.41	55.56	37.60
S2C	fastSAM	86.65	24.71	38.81	17.78
S2C	effi.SAM (vit-t)	90.57	39.04	47.65	42.92
S2C	effi.SAM (vit-s)	89.58	37.51	60.21	46.22
S2C	Dino-v2 (vit-b)	93.44	55.37	**61.17**	58.12
S2C + *IoU refine*	Dino-v2 (vit-b)	**93.75**	**57.68**	60.07	**58.85**

Table 2: Quantitative results of ablation study (CLCD).

(*Rec*) and F_1 score (Ding et al. 2024; Chen et al. 2023). *Pre* indicates the ratio of true positives among classified positives, while *Rec* is the measure of identifying true positives. F_1 is the harmonic mean of *Pre* and *Rec*.

Implementation Details

The training of S2C is performed using cropped images of 512×512 pixels over 10 epochs. The trained weights with highest accuracy on the validation set is saved for subsequent test evaluation. The training batch size depends on the backbone to fit in GPU memory, usually exceeding 12 in our implementation. The learning rate is initially set to 0.01, and is exponentially decayed with an factor of 1.5. The optimization algorithm is the Stochastic Gradient Descent with Nesterov momentum. The weighting parameters in Eq.2 are set to $\alpha = 0.2, \beta = 1$, while α can be adjusted across datasets to balance $\mathcal{L}_{tri}$ and $\mathcal{L}_{info}$ in training dynamics.

Ablation Study

Quantitative Evaluation. An ablation study is conducted through cumulative integration of the proposed methodologies, including the CTC, CSC, $\mathcal{L}_{spa}$ in Eq.6, and IoU matching and refinement (*IoU refine*). Given that the proposed S2C employs both VFM and CL, an intuitive baseline is to combine these two techniques. However, VFM alone is not capable for UCD, and there is no existing literature approach (to the best of our knowledge) that integrate CL with VFM for CD. Therefore, we implement these two baselines for comparison: i) applying CVA and clustering on the VFM-encoded semantic features for CD following the practice in (Zheng et al. 2024), and ii) conducting CS-based CL using the VFM features. These ablation experiments are evaluated using efficient-SAM (vit-t) (Xiong et al. 2024), an efficient variant of SAM (Kirillov et al. 2023).

The quantitative results are presented in Table 2. As indicated by the results of *eff.SAM + CVA*, direct change analysis on VFM features leads to suboptimal accuracy. By contrast, integrating VFM with SC-based CL demonstrates improved accuracy. The proposed CTC and CSC further surpasses SC-based CL paradigm. It is worth noting that the CTC alone outperforms both SC and CSC by a large margin, establishing it as an effective CL paradigm for UCD. Meanwhile, $\mathcal{L}_{spa}$ also exhibits notable effectiveness. Its addition on each CL paradigm results in an enhancement of average 2% in F_1. Adding CSC improves the robustness of CTC against temporal noise, leading to an increase of over 3% in F_1. The refining algorithm substantially enhances the precision of the results, resulting in an increase of 1.47% in F_1 and approximately 2% in OA.

S2C with different backbones. While S2C is introduced as a methodology that integrates CL and VFM, the core technique employed is a UCD framework, which is adaptable for other types of DNNs. Table 2 also presents an evaluation of S2C utilizing various different backbones, including a vanilla ResNet (He et al. 2016) and several other VFMs. Surprisingly, implementing S2C with a naive ResNet34 backbone still yields considerably high accuracy. This confirms its efficacy as a general framework for UCD.

Compared to vanilla DNNs, integrating VFMs in S2C greatly improves its Pre. Intuitively, the rich semantic contexts inherent to VFMs can facilitate the discrimination of semantic changes. The tested VFMs include fastSAM (Zhao et al. 2023), efficient SAM (Xiong et al. 2024) and Dino-v2(Oquab et al. 2023). The obtained metrics indicates that employing Dino-v2 results in the highest accuracy, with an advantage of approximately 12% compared to the other backbones. Therefore, Dino-v2 is selected as the S2C backbone in the subsequent experiments.

Qualitative Assessment. In Fig.4 we present some examples of the CD results obtained using different techniques. These results are selected in the 3 datasets covering different scenes, including (a) cropland, (b) countryside, (c) factories and (d) residential block. One can observe that using either CSC or CTC alone leads to much noise, while their collaborative employment greatly reduces false alarms. The grid sparsity generalization further removes much insignificant change representations and leads to smoothed segments. After the post-processing of *IoU refinement*, binary CD results matching the object boundaries are produced.

Comparative Experiments

We conduct acomparative experiments with the SOTA methods for UCD, including several non-DL methods based on difference analysis: CVA (Bruzzone and Prieto 2000), ISFA (Wu, Du, and Zhang 2014), DCVA (Saha, Bovolo, and Bruzzone 2019), DSFA (Du et al. 2019), KPCA-MNet (Wu et al. 2022) and SiROC (Kondmann et al. 2022). In addition, we also compare generative methods CDRL (Noh et al. 2022) and FCD-GAN(Wu, Du, and Zhang 2023), an augmentation-based method I3PE (Chen et al. 2023) and a recent VFM-based method AnyChange (Zheng et al. 2024). SAM-CD(Ding et al. 2024) is also included in the comparison, representing the SOTA accuracy of supervised CD.

In Table.3 the quantitative results are presented. Among the difference analysis-based methods, DCVA (Saha, Bovolo, and Bruzzone 2019) exhibits a notable advantage. It obtains the highest Rec on the CLCD and the highest Pre on the CLCD and SECOND. Among the methods presented in the recent literature, I3PE achieves a stable high accu-

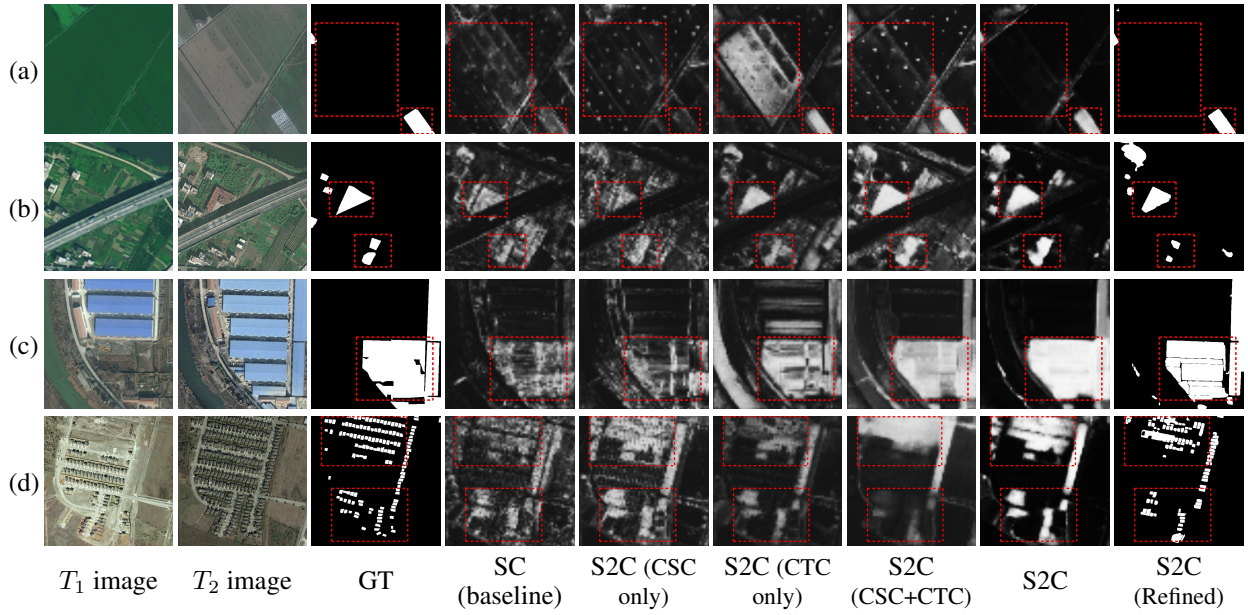

Figure 4: Example results obtained with different methods in the ablation study.

Methods	References	CLCD				SECOND				Levir			
		OA	Pre	Rec	F_1	OA	Pre	Rec	F_1	OA	Pre	Rec	F_1
SAM-CD	*TGRS 2024*	96.26	73.01	78.84	75.81	88.56	73.32	66.00	69.47	99.17	92.62	91.04	91.82
CVA	*TGRS 2000*	71.01	8.49	29.62	13.20	59.17	20.55	37.34	26.51	66.50	5.80	36.59	10.02
ISFA	*TGRS 2014*	74.37	8.60	25.39	12.85	60.23	20.13	34.24	25.35	69.32	6.03	34.45	10.27
DCVA	*TGRS 2019*	53.91	11.35	**76.26**	19.76	55.51	25.63	66.07	36.93	48.28	7.20	76.94	13.16
DSFA	*TGRS 2019*	52.09	8.53	55.94	14.80	48.10	19.06	50.28	27.65	60.29	5.99	46.22	10.60
KPCA	*TCYB 2022*	53.47	13.84	44.52	21.11	54.69	20.44	44.87	28.09	54.97	5.61	49.52	10.08
CDRL	*CVPR 2022*	64.74	7.75	34.27	12.64	62.65	37.78	22.90	28.51	62.59	5.32	37.74	9.32
SiROC	*TGRS 2022*	82.38	18.93	41.65	26.03	69.80	27.92	33.59	30.49	64.82	7.38	51.14	12.90
I3PE	*ISPRS 2023*	91.12	32.42	17.79	22.97	74.09	34.53	35.03	34.78	91.03	17.39	20.31	18.74
FCD-GAN	*TPAMI 2023*	83.93	22.06	45.79	29.77	68.36	29.47	43.41	35.11	83.42	8.87	24.29	12.99
AnyChange	*NeurIPS 2024*	-	-	-	-	-	30.5	**83.2**	44.6	-	13.3	**85.0**	23.0
S2C-Dinov2	proposed	93.95	59.04	61.24	60.12	**84.55**	**67.15**	42.41	51.99	89.28	29.42	78.88	42.86
S2C + *IoU Refine*	proposed	**94.46**	**63.82**	59.12	**61.38**	84.45	64.91	46.02	**53.86**	**92.84**	**34.85**	70.69	**46.69**

Table 3: Quantitative accuracy (%) evaluation of the proposed S2C and SOTA UCD methods on various benchmark datasets.

racy in the three datasets. AnyChange (Zheng et al. 2024), a training-free approach of leveraging VFM for CD, achieves the highest Rec on two datasets. However, it obtains a relatively low Pre, suggesting a considerable percentage of false alarms present in its results.

The proposed S2C yields substantial and consistent ccuracy improvements over the SOTA. The preliminary predictions generated by S2C lead to a significant enhancement exceeding 30% on the CLCD dataset and approaching 20% on the Levir dataset. The subsequent post-processing stage, which integrates IoU refinement, further optimizes the trade-off between Pre and Rec. The improvements of S2C over the SOTA methods, after refinement, are quantified as 31%, 9%, and 23% in F_1 for the respective datasets. The proposed S2C also significantly reduces the accuracy gap with fully supervised methods, as shown by reductions of 14% and 16% in F_1 on the CLCD and SECOND, respectively.

Conclusions

This study formulates a CL framework to explicitly model unsupervised learning of semantic changes in VHR RS images. To address the challenges including spectral variations and spatial misalignment in CD of VHR RS images, two consistency-regularized CL paradigms, i.e., the CSC and CTC, are developed. Notably, the CTC paradigm presents an innovative multi-temporal triplet learning strategy to address the existing gap in explicit difference learning. In addition, several novel techniques are developed to translate the VFM semantics into CD results, including grid sparsity regularization, negative cosine embedding of changes, and an IoU refinement algorithm. The developed CL framework, S2C, exhibits notable superiority over SOTA methods, as evaluated on three CD benchmark datasets. Future studies are recommended to extend the S2C to more challenging UCD tasks, such as semantic CD and multimodal CD.

Acknowledgments

This work is funded by the National Natural Science Foundation of China under Grant No.42201443 and No.42571458. It is also funded by the Henan Provincial Key Technologies R & D Program under Grant 242102211047.

References

Bandara, W. G. C.; and Patel, V. M. 2022. Revisiting Consistency Regularization for Semi-supervised Change Detection in Remote Sensing Images.

Bandara, W. G. C.; and Patel, V. M. 2023. Deep Metric Learning for Unsupervised Remote Sensing Change Detection. *arXiv preprint arXiv:2303.09536*.

Bruzzone, L.; and Prieto, D. F. 2000. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 38(3): 1171–1182.

Chen, H.; Qi, Z.; and Shi, Z. 2021. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

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, H.; Song, J.; Wu, C.; Du, B.; and Yokoya, N. 2023. Exchange means change: An unsupervised single-temporal change detection framework based on intra-and inter-image patch exchange. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206: 87–105.

Chen, H.; Song, J.; and Yokoya, N. 2024. Change Detection Between Optical Remote Sensing Imagery and Map Data via Segment Anything Model (SAM). *arXiv preprint arXiv:2401.09019*.

Chen, H.; Yokoya, N.; Wu, C.; and Du, B. 2022. Unsupervised Multimodal Change Detection Based on Structural Relationship Graph Representation Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Chen, Y.; and Bruzzone, L. 2021. Self-Supervised Change Detection in Multiview Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12.

Ding, L.; Hong, D.; Zhao, M.; Chen, H.; Li, C.; Deng, J.; Yokoya, N.; Bruzzone, L.; and Chanussot, J. 2025. A Survey of Sample-Efficient Deep Learning for Change Detection in Remote Sensing: Tasks, strategies, and challenges. *IEEE Geoscience and Remote Sensing Magazine*.

Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Yang, K.; and Bruzzone, L. 2024. Adapting Segment Anything Model for Change Detection in HR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–11.

Dong, S.; Wang, L.; Du, B.; and Meng, X. 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208: 53–69.

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.

Gao, F.; Dong, J.; Li, B.; and Xu, Q. 2016. Automatic Change Detection in Synthetic Aperture Radar Images Based on PCANet. *IEEE Geoscience and Remote Sensing Letters*, 13(12): 1792–1796.

Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. 2024. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27672–27683.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; Plaza, A.; Gamba, P.; Benediktsson, J. A.; and Chanussot, J. 2024. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI:10.1109/TPAMI.2024.3362475.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Ji, W.; Li, J.; Bi, Q.; Liu, T.; Li, W.; and Cheng, L. 2024. Segment anything is not always perfect: An investigation of sam on different real-world applications.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Kondmann, L.; Toker, A.; Saha, S.; Schölkopf, B.; Leal-Taixé, L.; and Zhu, X. X. 2022. Spatial Context Awareness for Unsupervised Change Detection in Optical Satellite Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Li, K.; Cao, X.; Deng, Y.; Pang, C.; Xin, Z.; Meng, D.; and Wang, Z. 2025. DynamicEarth: How Far are We from Open-Vocabulary Change Detection? *arXiv preprint arXiv:2501.12931*.

Liu, M.; Chai, Z.; Deng, H.; and Liu, R. 2022. A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 4297–4306.

Liu, T.; Zhang, M.; Gong, M.; Zhang, Q.; Jiang, F.; Zheng, H.; and Lu, D. 2025. Commonality feature representation learning for unsupervised multimodal change detection. *IEEE Transactions on Image Processing*.

Mall, U.; Hariharan, B.; and Bala, K. 2023. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5261–5270.

Noh, H.; Ju, J.; Seo, M.; Park, J.; and Choi, D.-G. 2022. Unsupervised change detection based on image reconstruction loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1352–1361.

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*.

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.

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.

Saha, S.; Bovolo, F.; and Bruzzone, L. 2020. Building change detection in VHR SAR images via unsupervised deep transcoding. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3): 1917–1929.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Wang, L.; Zhang, M.; and Shi, W. 2023. CS-WSCDNet: Class Activation Mapping and Segment Anything Model-Based Framework for Weakly Supervised Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Wu, C.; Chen, H.; Du, B.; and Zhang, L. 2022. Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network. *IEEE Transactions on Cybernetics*, 52(11): 12084–12098.

Wu, C.; Du, B.; and Zhang, L. 2014. Slow Feature Analysis for Change Detection in Multispectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(5): 2858–2874.

Wu, C.; Du, B.; and Zhang, L. 2023. Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9774–9788.

Xiong, Y.; Varadarajan, B.; Wu, L.; Xiang, X.; Xiao, F.; Zhu, C.; Dai, X.; Wang, D.; Sun, F.; Iandola, F.; et al. 2024. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16111–16121.

Yang, K.; Xia, G.-S.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; and Zhang, L. 2022. Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7236–7246.

Zhao, M.; Hu, X.; Zhang, L.; Meng, Q.; Chen, Y.; and Bruzzone, L. 2024. Beyond pixel-level annotation: Exploring self-supervised learning for change detection with image-level supervision. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

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. *Advances in Neural Information Processing Systems*, 37: 81204–81224.