Learning Semantic-Robust Change Detection via Semantic-Invariant Self-Distillation

Jiuhe Qu[1]*, Yingping Liang[1]*, and Ying Fu[1]†

Beijing Institute of Technology, Beijing, China
{qujiuhe,liangyingping,fuying}@bit.edu.cn

Abstract. Change detection aims to identify semantic changes between remote sensing images. However, features from models are easily disturbed by non-semantic variations, such as illumination, shadows, and atmospheric changes, leading to false alarms and limited generalization in real-world scenarios. In this paper, we propose **SCDistill**, a framework for learning semantic-robust change detection via semantic-invariant self-distillation. First, to strengthen semantic consistency, we introduce a semantic-invariant self-distillation strategy that learns semantic robustness from perturbed yet semantically consistent data, empowering the change detector to extract disturbance-resistant features and achieve more reliable and accurate semantic change identification. Second, to expand paired data with non-semantic variations, we design a diffusion-based perturbation simulation pipeline that synthesizes complex environmental changes, enabling the model to explicitly learn to distinguish semantic changes from appearance-level fluctuations and reduce false alarms caused by non-semantic disturbances. These components promote robustness from data and representation perspectives, leading to synergistic performance gains. Extensive experiments demonstrate that SCDistill achieves state-of-the-art performance on multiple semantic change detection benchmarks and exhibits strong generalization to binary change detection and change captioning tasks. Code is accessible at `https://github.com/elecreak/SCDistill`.

Keywords: Change detection · Self-distillation · Representation learning · Remote sensing

1 Introduction

Change detection (CD) plays a crucial role in remote sensing data analysis, aiming to identify semantic changes between bi-temporal images. This capability facilitates numerous applications, including development monitoring [42], land-cover mapping [29,53], and environmental surveillance [23]. With the development of deep learning [35, 57, 66, 69], recent works [7, 11, 33, 77] have achieved accuracy gains by modeling semantic segmentation and temporal differencing.

* Equal contribution. † Corresponding author.

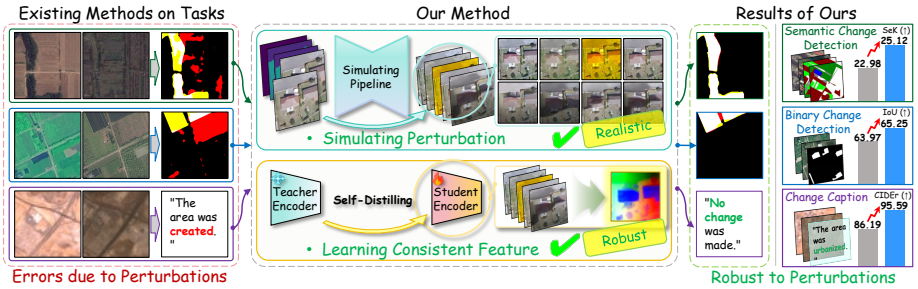

Fig. 1: Existing methods rely on limited data collections that lack non-semantic variations and general-purpose encoders that are sensitive to appearance changes. In contrast, **SCDistill** leverages simulated non-semantic perturbations and a semantic-invariant self-distilled model, achieving consistent improvements across **three representative tasks**. Regions highlighted in **Red** indicate incorrect detections.

Despite this progress, deep learning-based CD remains vulnerable to non-semantic variations—such as illumination shifts, shadows, atmospheric effects, and seasonal transitions—that frequently occur in real-world remote sensing imagery. These disturbances often change surface appearance without altering underlying semantics, yet existing models [7, 16, 33] may mistakenly interpret them as semantic changes, leading to false alarms and degraded robustness. A key reason is that most CD pipelines heavily rely on vision encoders pre-trained on single-view classification or segmentation tasks [7, 16, 33], where representations are typically appearance-sensitive. When such disturbance-sensitive features are propagated into temporal differencing and detection heads, predictions become unstable under real-world environmental fluctuations.

To address this issue, we propose **SCDistill**, a framework for **semantic-robust change detection** via **semantic-invariant self-distillation**. The core idea is to explicitly distill semantic invariance into the CD model so that it produces consistent semantic representations and change predictions under diverse appearance conditions. Specifically, we introduce a **semantic-invariant self-distillation** strategy, where a vision foundation model encoder (VFME) serves as a semantic teacher and guides the student to learn disturbance-resistant representations. Given a clean bi-temporal pair and its appearance-perturbed yet semantically consistent counterpart, the student is trained to align its predictions across conditions to the semantic guidance from teacher. This cross-condition distillation enforces semantic consistency while suppressing appearance-sensitive noise, enabling the student to inherit stronger semantic abstraction and achieve more reliable change detection in complex real-world scenarios.

A practical challenge in applying such semantic-invariant self-distillation is the lack of paired training samples exhibiting semantic-preserving appearance variations. Most existing synthetic CD datasets generate bi-temporal pairs via label-guided synthesis or semantic region replacement, while largely overlooking real-world non-semantic perturbations that occur globally and especially in

unchanged regions without label variation [4, 71]. Consequently, models trained on these idealized datasets tend to overfit to clean scenes without non-semantic appearance variations and perfectly aligned conditions, and the distillation objective cannot be sufficiently exercised across diverse environmental conditions.

Motivated by this, from the perspective of data, we design a diffusion-based **perturbation simulation** pipeline to enrich synthetic CD datasets with realistic, semantic-invariant environmental variations, including illumination shifts, atmospheric and weather changes. This provides the necessary cross-condition supervision for semantic-invariant distillation, allowing the model to explicitly learn to distinguish true semantic changes from appearance-level variations and thereby reduce false alarms. Together, semantic-invariant self-distillation and diffusion-based perturbation simulation jointly improve robustness from representation and data perspectives, leading to more stable and accurate semantic change detection under real-world disturbances.

Extensive experiments demonstrate that SCDistill achieves state-of-the-art performance on multiple semantic change detection benchmarks and generalizes effectively to related tasks, including binary change detection and change captioning, showcasing its strong robustness in real-world scenarios and generalization applicability in related tasks. The proposed components promote robustness from data and representation perspectives, leading to synergistic performance gains. In summary, our main contributions are as follows:

- We introduce **SCDistill**, a framework for semantic-robust change detection that mitigates non-semantic disturbances through diffusion-based perturbation simulation and semantic-invariant self-distillation.
- We design a **semantic-invariant self-distillation** strategy that learns semantic robustness from perturbed yet semantic-invariant data by self-distillation, enabling disturbance-resistant feature representations for accurate semantic change identification.
- We propose a diffusion-based **perturbation simulation** pipeline that synthesizes complex environmental variations, enabling the model to explicitly learn to distinguish semantic changes from appearance-level fluctuations.

2 Related Work

Change Detection Methods. CNN-based methods are fundamental to deep visual tasks [17, 24, 68, 70] and have been widely used in change detection, with existing designs mainly categorized by bitemporal image integration strategies. Input-level concatenation, such as stacking and differencing [58], or principal component analysis [1] of bitemporal images before feeding them into a single encoder-decoder, such as Fully Convolutional Network (FCN) [43] and U-Net [51], for change prediction. Dual-branch encoders with shared weights process each temporal image, and features are fused by concatenation and subtraction in intermediate layers [12,21,30,34]. Other works extend the dual-branch net by incorporating attention mechanisms, such as channel [27, 59], spatial [2], or spatio-temporal [44, 61], to refine feature fusion and emphasize change-relevant

regions. However, they remain sensitive to non-semantic disturbances such as illumination and atmospheric changes, often leading to false detections.

Synthetic Datasets for Change Detection. Synthetic datasets are widely used to alleviate annotation scarcity [31,32,36,37,67,78], and have recently been used in change detection. Early graphics-based approaches, such as AICD [5] and SyntheWorld [56] render 3D scenes but suffer from limited photorealism and diversity. Hybrid pipelines, such as IAug [8] and Changen [72] insert segmented objects into real images using GANs. Diffusion-driven methods like Changen2 [71], HySCDG [4], and ChangeDiff [64] synthesize change sequences via generation or text-guided inpainting. These synthetic datasets emphasize semantic transitions but overlook real-world variations like illumination and shadow changes in unchanged regions, causing unreliable supervision.

Vision Foundation Model Encoders. Deep learning models pre-trained on large-scale datasets serve as vision fundamental feature extract models for downstream vision tasks. Widely adopted architectures include ResNet [25] trained on ImageNet [15] and fully convolutional networks (FCN) [43] for dense prediction. Recently, self-supervised Vision Foundation Model like DINOv3 [55], composed of vision transformers, has emerged as a powerful alternative, capturing rich semantic representations without manual annotations. Although vision foundation models provide rich semantic priors, their features often lack sufficient discrimination between semantic and appearance-level variations, which can lead to ambiguous representations and unreliable guidance for downstream decoding.

3 Method

In this section, we first introduce the formulation and motivation of our approach. Then, we present the perturbation simulation, where a pipeline synthesizes non-semantic disturbances to model realistic appearance variations. Next, we describe the proposed semantic-invariant self-distillation strategy that learns disturbance-resistant semantic representations for robust semantic change detection, as illustrated in Figure 2. Finally, we provide the learning details.

3.1 Formulation and Motivation

Change detection (CD) aims to identify differences between bi-temporal images $\{\mathbf{I}_1, \mathbf{I}_2\}$ of the same scene. Depending on the output form, CD can be performed at different levels, including semantic CD, binary CD, and change captioning. All these tasks share a common encoder for feature $\mathbf{F}$ extraction, with task-specific decoders producing change maps or textual descriptions $\mathbf{C}$:

$$\mathbf{F} = \mathrm{Enc}(\{\mathbf{I}_1, \mathbf{I}_2\}), \quad \mathbf{C} = \mathrm{Dec}(\mathbf{F}). \tag{1}$$

Since large-scale bi-temporal datasets with pixel-level annotations are difficult to collect, some recent works synthesize data by generating semantic layouts $\mathbf{S}$ and producing corresponding images through guided rendering:

$$\{\mathbf{I}_1, \mathbf{I}_2, \mathbf{L}\} = \mathrm{Synthesize}(\mathbf{S}; \Phi), \tag{2}$$

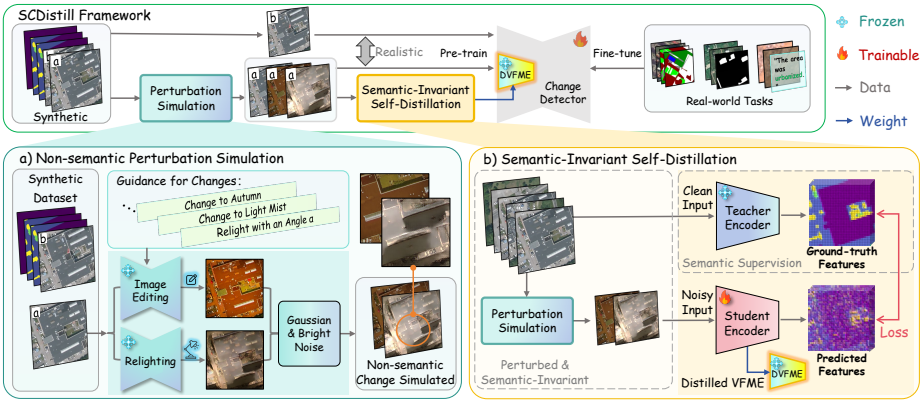

Fig. 2: Overview of the proposed SCDistill framework. a) The diffusion-based perturbation simulation synthesizes environmental variations to explicitly model real-world appearance disturbances, enhancing robustness against non-semantic changes. b) Meanwhile, the vision foundation model encoder obtain semantic supervision from the clean image in semantic-invariant self-distillation, guiding the disturbed student to learn disturbance-resistant representations for accurate semantic change detection.

where Φ denotes the synthesis process that generates bi-temporal images $\{\mathbf{I}_1, \mathbf{I}_2\}$ and their semantic labels $\mathbf{L}$.

However, there are still two main challenges. First, existing synthetic datasets primarily emphasize semantic region changes, while neglecting realistic non-semantic variations in unchanged regions, such as illumination shifts, shadows, and atmospheric effects, commonly observed in real-world imagery. Models trained on such data tend to overfit to idealized semantic transitions and struggle to distinguish true semantic changes from appearance-level disturbances. To address this, we propose a **perturbation simulation** strategy that explicitly models non-semantic variations using a diffusion-based renderer, converting existing synthetic datasets into more realistic bi-temporal scenes. This exposes the model to physically plausible disturbances during training, fostering more robust visual reasoning under real-world conditions.

Moreover, even with realistic perturbations, standard visual encoders pre-trained on single-view tasks remain sensitive to appearance fluctuations, producing inconsistent semantic representations that undermine downstream CD. We introduce a **semantic-invariant self-distillation** framework that transfers high-level semantic knowledge from a clean-image teacher to a student trained on perturbed counterparts. This cross-condition distillation enforces semantic consistency while filtering out non-semantic noise, yielding disturbance-robust features that enable reliable and accurate change identification. Together with perturbation simulation, this approach addresses both data and feature level limitations, forming a framework for semantic-robust change detection.

3.2 Non-semantic Perturbation Simulation

Synthetic change detection datasets alleviate the scarcity of annotated data, yet they usually fail to simulate the non-semantic changes that are widely present in real scenarios. Consequently, models trained on synthetic data often show poor robustness to such disturbances, which adversely affects their fine-tuning performance. To address this limitation, we augment synthetic datasets with complex environmental non-semantic variations that imitate realistic variations between two temporally captured images of the same scene. Specifically, we analyze real-world bi-temporal change detection datasets and identify typical non-semantic change types that frequently lead to false detections in existing methods. These include changes in illumination and shadow caused by sunlight direction and weather variations, which significantly influence model predictions.

Data Source and Notation. To enable the simulated non-semantic perturbation samples to be learned by the CD model, we start from an existing synthetic change detection dataset that provides semantically edited image pairs. This dataset is generated by modifying specific regions of real remote sensing images to create semantic changes. Given a sample triplet $\{\mathbf{X}_0, \mathbf{X}_1, \mathbf{L}\}$, where $\mathbf{X}_0$ and $\mathbf{X}_1$ are pre- and post-change images and $\mathbf{L}$ is the semantic change labels, since non-semantic disturbances between bi-temporal images can be introduced by altering only one image of each pair, we apply these perturbations to the pre-change image $\mathbf{X}_0$, which can be formulated as:

$$\tilde{\mathbf{X}}_0 = \text{Perturb}(\mathbf{X}_0), \quad \tilde{\mathbf{X}}_1 = \mathbf{X}_1, \quad \tilde{\mathbf{L}} = \mathbf{L}. \tag{3}$$

This preserves the semantic editing applied in the synthetic dataset while injecting realistic appearance variations between the bi-temporal image pairs.

Non-semantic Variations Simulation. Different categories of non-semantic changes are simulated using different models. For illumination and shadow variations induced by solar angle changes, we employ a scene illumination modification model such as DIR [63] to simulate lighting conditions. For weather and atmospheric variations, we adopt an image editing model such as InstructPix2Pix [6], guided by multiple pre-defined textual prompts p describing various non-semantic changes. Through these processes, multiple perturbations simulated versions of $\mathbf{X}_0$ are produced, which can be formulated as:

$$\tilde{\mathbf{X}}_0^{(i)} = \{\text{Relighting}(\mathbf{X}_0), \text{Editing}(\mathbf{X}_0, p_i)\}, \quad i = 1, \ldots, N. \tag{4}$$

This results in a set of noise-injected images $\{\tilde{\mathbf{X}}_0^{(1)}, \tilde{\mathbf{X}}_0^{(2)}, \ldots\}$ that reflect a wide range of realistic non-semantic changes. Combined with the original $\mathbf{X}_1$ and label $\mathbf{L}$, this forms a new large-scale dataset that simulates both semantic and non-semantic changes, bridging the gap between synthetic and real-world data.

Compared with GAN-based [60] methods or only making modifications at the pixel level, such as Gaussian blur, random noise, and color jitter, the diffusion-based simulation offers controllable and diverse perturbations with high visual fidelity, ensuring the generated non-semantic variations remain consistent with real-world conditions. This ensures the dataset bridges the domain gap and preserves semantic integrity crucial for subsequent distillation.

3.3 Semantic-Invariant Self-Distillation

While the perturbation simulation exposes the model to realistic non-semantic variations, it does not directly guarantee semantic robustness at the feature level. Therefore, we introduce a semantic-invariant self-distillation strategy to explicitly inject such robustness into the encoder representations. Most existing semantic change detection methods employ general-purpose visual encoders pretrained on single-image recognition tasks, such as the ResNet [25] used in Bi-SRNet [16] or the X3D [20] backbone in Change3D [77]. However, these encoders, trained on single-view datasets, are insufficient to capture semantic consistency across complex real-world variations, resulting in degraded robustness in semantic feature extraction. To address this limitation, we introduce a semantic-invariant encoder that learns to extract noise-resistant representations by distilling knowledge from diverse non-semantic variations synthesized from large-scale data.

Noise-Augmented Pairs Construction. To enable this distillation, we prepare a dataset consisting of image pairs with simulated non-semantic perturbations while maintaining semantic invariance. For each pair, the first image is sampled from a single-temporal remote sensing dataset, and the second is generated by applying the non-semantic variation simulation pipeline described in perturbation simulation, which means the perturbation-simulated dataset can be used as this dataset. Additionally, random low-level perturbations such as Gaussian blur, grayscale conversion, and brightness and saturation shifts are applied to further increase pixel-level non-semantic discrepancies. As a result, each image pair shares identical semantic content but differs in appearance due to realistic non-semantic variations.

Robust Representations Learning. We employ a distillation-based training strategy to transfer semantic invariance into the pre-trained vision foundation model encoder (VFME). Given a single-temporal image $\mathbf{X}$, we first generate its noise-augmented version $\tilde{\mathbf{X}} = \text{Perturb}(\mathbf{X})$. We use the VFME pretrained on general tasks as a frozen teacher and to initialize the student. The teacher E_T extracts a clean semantic representation $\mathbf{F}_T$ from the original image, which serves as the target for the student model E_s trained on the noisy input affected by non-semantic perturbations, which can be formulated as:

$$\mathbf{F}_T = \text{E}_T(\mathbf{X}), \quad \mathbf{F}_S = \text{E}_S(\tilde{\mathbf{X}}), \tag{5}$$

where $\mathbf{F}_S$ and $\mathbf{F}_T$ are features from student and teacher. We then minimize the discrepancy between $\mathbf{F}_S$ and $\mathbf{F}_T$ using L2 regression loss in distilling.

Through training on diverse noise conditions, the distilled VFME, as the student encoder, learns to produce semantic-invariant semantic features that remain consistent with clean outputs from the teacher model. Importantly, this training requires no change labels or semantic segmentation supervision, allowing it to scale across large unlabeled datasets. The resulting encoder is then used in downstream semantic change detection models to enhance robustness against real-world appearance shifts. Through this process, the student encoder internalizes semantic invariance, serving as a robust foundation for subsequent change detection models under diverse environmental conditions. The teacher

Table 1: Performance comparison of training on multiple semantic CD benchmarks with the state-of-the-art method. The best results are represented in bold, and the second-best results are indicated with an underline. All results are described aspercentages (%).

Method	Backbone	SECOND [62]				HRSCD [14]			
		Fscd	mIoU	OA	SeK	Fscd	mIoU	OA	SeK
HRSCD-S4 [14] [CVIU'2019]	ResNet	58.21	71.15	86.62	18.80	69.39	67.79	81.32	22.50
SCDNet [49] [JPRS'2022]	ResNet	60.01	70.97	87.40	19.73	70.80	67.43	81.46	23.61
ChangeMask [74] [IJAEOG'2021]	EfficientNet-B0	59.74	71.46	86.93	19.50	70.59	67.56	81.65	23.43
SSCD-L [16] [TGRS'2022]	ResNet	61.22	72.60	87.19	21.86	69.69	66.21	81.05	21.88
Bi-SRNet [16] [TGRS'2022]	ResNet	61.85	72.08	87.20	21.36	71.72	67.83	82.06	24.59
MTSCD [13] [IJAEOG'2023]	ResNet	60.23	71.68	87.04	20.57	67.56	63.71	72.06	9.03
JFRNet [7] [TGRS'2024]	ResNet	62.63	72.82	87.10	22.56	66.65	64.65	76.20	18.78
DEFO [33] [TGRS'2024]	ResNet	61.18	72.39	87.01	21.08	66.49	65.67	81.36	19.20
ChangeMamba [11] [TGRS'2024]	Mamba-based	61.59	72.32	87.84	20.30	70.04	67.93	82.68	24.50
Changen2 [71] [TPAMI'2024]	ViT	62.79	72.39	87.67	22.09	73.03	68.73	82.82	26.25
Change3D [77] [CVPR'2025]	VideoEncoder	62.83	72.95	87.42	22.98	73.29	68.67	82.57	26.85
SCDistill (Ours)	Distilled VFME	**65.64**	**74.01**	**88.59**	**25.12**	**74.29**	**70.28**	**83.61**	**28.38**

only receives clean inputs and provides semantic-invariant supervision as targets of distillation. So the robustness of the student is not inherited from a robust teacher, but is mainly induced by the semantic-invariant distillation process.

3.4 Learning Details

To implement the perturbation simulation, we use the DIR [63] model to simulate illumination changes, where the azimuth angle is uniformly from 0° to 360° and the elevation angle from 20° to 70°. When employing InstructPix2Pix [6] for weather and atmospheric editing, a randomly selected pre-defined prompt is used for each input image to direct the modification process.

To construct a change detector guided by feature representations with robustness against non-semantic interference, we adopt the video encoder–universal decoder from Change3D [77] and integrate the disturbance-robust semantic encoder distilled from the VFME into the encoder branch. This self-distilled encoder provides semantically stable features $\mathbf{F}_0$ and $\mathbf{F}_1$ for the pre-change and post-change images $(\mathbf{I}_0, \mathbf{I}_1)$, which are fused with the shallow X3D [20] representations to enhance resilience against non-semantic perturbations and guide the deeper layers to focus on meaningful semantic change regions while suppressing irrelevant appearance-level fluctuations.

During the training of our CD network, the model is pre-trained on the diffusion-augmented synthetic dataset to learn invariance and fine-tuned on real-world benchmarks. Following [77], we adopt task-specific joint losses as follows cross-entropy, Dice [46], and cosine similarity [16] for semantic CD; cross-entropy and Dice for binary CD; and cross-entropy for change captioning.

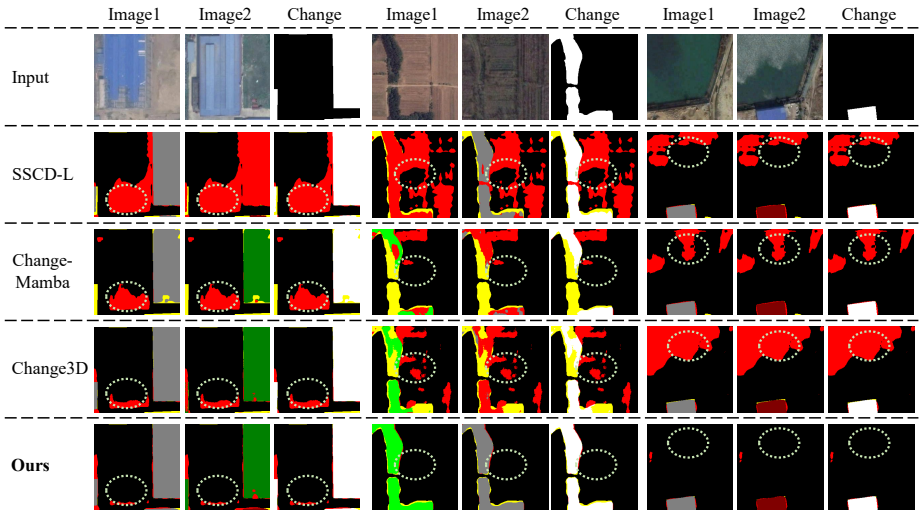

Fig. 3: Qualitative comparison with representative state-of-the-art methods on the SECOND [62] dataset. Black represents non-change, **Red** and **Yellow** indicate **False** and **Missed** detections, while white and other colors denote correct change regions and their semantic categories respectively.

4 Experiments

In this section, we first introduce the datasets and evaluation metrics for experiments. Then, performance comparisons are conducted with the other methods. Finally, ablations, visualized qualitative results, and discussions are provided to confirm the effectiveness of the main components.

4.1 Experimental Settings

Datasets. To provide abundant data for pretraining and to cover a wide range of semantic change types, we adopt the large-scale synthetic dataset FSC-180k [4], which includes 20 categories of semantic change simulated from real remote sensing images across diverse land-cover types. For evaluation, multiple real-world benchmarks are utilized, including HRSCD [14] and SECOND [62], which contain diverse non-semantic variations such as illumination, shadow, and atmospheric changes commonly observed in real scenarios. For binary CD and change captioning tasks, we further employ CLCD [40] and DUBAI-CC [26], as their feature are diverse land-cover transitions and higher semantic complexity, posing greater challenges to model robustness and generalization. During both training and testing, all datasets are processed to a spatial resolution of 256×256. Specifically, for CD datasets, bi-temporal image pairs and corresponding labels are synchronously cropped into non-overlapping patches, while for change captioning datasets, image resolutions are directly resized to the same scale.

Table 2: Performance comparison of training on binary CD task benchmarks.

Method	CLCD [40]			LEVIR-CD [10]			WHU-CD [28]		
	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
DTCDSCN [42] [GRSL'2020]	57.47	40.81	94.59	87.43	77.67	98.75	79.92	66.56	98.05
SNUNet [19] [GRSL'2021]	60.82	43.63	94.90	88.16	78.83	98.82	83.22	71.26	98.44
ChangeFormer [3] [IGRSS'2022]	61.31	44.29	94.98	90.40	82.48	99.04	87.39	77.61	99.11
BIT [9] [TGRS'2021]	59.93	42.12	94.77	89.31	80.68	98.92	83.98	72.39	98.52
ICIFNet [22] [TGRS'2022]	68.66	52.27	95.77	89.96	81.75	98.99	88.32	79.24	98.96
Changer [18] [TGRS'2023]	67.07	50.46	95.69	90.70	82.99	99.08	89.18	80.48	99.17
DMINet [21] [TGRS'2023]	67.24	50.65	95.21	90.71	82.99	99.07	88.69	79.68	98.97
GASNet [65] [ISPRS'2023]	63.84	46.89	94.01	90.52	83.48	99.07	91.75	84.76	99.34
AMTNet [41] [ISPRS'2023]	75.10	60.12	96.45	90.76	83.08	98.96	92.27	85.64	99.32
Self-Pair [52] [WACV'2023]	75.49	60.63	96.42	90.83	83.20	99.07	91.59	84.94	99.03
Changen [73] [ICCV'2023]	76.96	62.88	96.31	91.12	83.69	99.11	92.47	87.10	98.95
EATDer [45] [TGRS'2024]	72.01	56.19	96.11	91.20	83.80	98.75	90.01	81.97	98.58
AnyChange [75] [NeurIPS'2024]	75.93	61.20	96.51	90.65	82.91	99.06	92.87	86.02	99.32
Changen2 [71] [TPAMI'2024]	78.00	63.84	96.89	91.79	84.62	99.16	94.62	89.75	99.43
Change3D [77] [CVPR'2025]	78.03	63.97	96.87	91.82	84.87	**99.17**	94.56	89.69	99.57
SCDistill (Ours)	**78.97**	**65.25**	**96.93**	**91.88**	**85.04**	99.16	**94.79**	**89.84**	**99.60**

Evaluation Metrics. For semantic CD, we report Fscd (F1 score on changed regions), mean Intersection over Union (mIoU), Overall Accuracy (OA), and Separate Kappa (SeK). For binary CD, we use standard metrics including pixel-wise F1 score, IoU, and Overall Accuracy (OA). For change captioning, we use BLEU-1 and BLEU-2 for their focus on lexical and phrasal precision, which is important for CD tasks, along with METEOR, ROUGE-L, and CIDEr-D, as adopted in change captioning literature [38].

Training Parameters. For the self-distillation of the semantic encoder, we employ the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.05, using a batch size of eight per GPU. For the pretraining and fine-tuning of the change detection model, we adopt the Adam optimizer with $\beta = (0.9, 0.99)$, a weight decay of 2×10^{-4}, and an initial learning rate of 2×10^{-4} that decays following the schedule $(1 - \text{iter}_{\text{curr}}/\text{iter}_{\text{max}})^{\alpha} \times \text{lr}$, where α is set to 0.9. All models are trained on 256×256 image patches with data augmentation techniques including random resizing, flipping, and cropping. Distillation takes 5 hours, pretraining takes 20 hours as a one-time cost. The simulation pipeline takes only 8 hours on 1 NVIDIA RTX 3090, with 0.7 seconds per image. The fine-tuning deployment takes 8 hours, as the pretrained model and self-distilled encoder trained by the perturbation simulated data can be used across tasks directly without cost.

4.2 Main Results

Quantitative Results on Semantic CD. As shown in Table 1, we compare our method with representative semantic CD approaches under standard supervised settings, including classical models (HRSCD-S4 [14], SCDNet [49]), recent convolution-based frameworks (SSCD-L [16], Bi-SRNet [16], DEFO [33],

MTSCD [13], JFRNet [7]), and advanced designs such as ChangeMask [74] (EfficientNet-B0 backbone), ChangeMamba [11] (Mamba-based), Changen2 [71] (pretrained with generative change supervision), and Change3D [77] (X3D-based video encoder [20]). In contrast to these methods, our method integrates the semantic-invariant self-distillation with non-semantic perturbations simulation, enabling the model to suppress non-semantic feature noise and maintain semantic consistency. This design yields consistent improvements across benchmarks, with particularly notable gains in semantic-related metrics; for example, on SECOND, our method surpasses the state-of-the-art method by 9.31% in SeK and 4.47% in Fscd, confirming its robustness under complex scenarios.

Qualitative Results on Semantic CD. Figure 3 presents qualitative comparisons between our method and three representative baselines, selected to cover diverse architectural paradigms: SSCD-L [16] (ResNet-based), ChangeMamba [11] (Mamba-based), and Change3D [77] (video-encoder-based, current state-of-the-art). These models represent different backbone families and performance tiers, providing a comprehensive visual comparison of semantic change detection quality. Our method achieves more accurate change localization and semantic classification in complex real-world scenes, while other methods tend to misclassify areas affected by non-semantic environmental variations such as shadows, illumination, or weather shifts. These results visually confirm that SCDistill effectively suppresses irrelevant appearance changes and maintains semantic consistency, enabling robust deployment in challenging environments.

Generalization to Binary CD. As shown in Table 2, we evaluate the generalization ability of our SCDistill framework from semantic CD to binary CD under standard supervised settings. We compare our approach with representative methods, including DTCDSCN [42], SNUNet [19], ChangeFormer [3], BIT [9], ICIFNet [22], Changer [18], DMINet [21], GASNet [65], AMTNet [41], Self-Pair [52], Changen [73], EATDer [45], AnyChange [75], Changen2 [71], and Change3D [77]. Across all challenging datasets, our method consistently outperforms existing approaches, achieving state-of-the-art performance. In particular, we observe significant gains in accuracy-sensitive metrics (F1 and IoU), demonstrating that the semantic-invariant self-distillation and synthetic perturbation pretraining effectively enhance the robustness to non-semantic variations of the encoder and improve its generalization capability across change detection tasks.

Generalization to Change Captioning. As shown in Table 3, we further assess the generalization of our SCDistill framework to the change captioning task. We compare against recent advanced approaches, including DUDA [48], MCCFormer [50], RSICC [38], SEN [39], PromptCC [76], and Change3D [77]. Our method achieves superior performance across all benchmarks, especially on semantically sensitive metrics such as METEOR, ROUGE-L, and CIDEr. These improvements indicate that our semantic-invariant self-distillation strategy enables the model to maintain semantic consistency under various appearance perturbations, thereby producing more precise and semantically faithful textual descriptions of changes.

Table 3: Performance comparison of training on change captioning task benchmark DUBAI-CC [26] with state-of-the-art methods.

Method	BLEU-1	BLEU-2	METEOR	ROUGE	CIDEr
DUDA [48] [ICCV'2019]	58.82	43.59	22.05	48.34	62.78
MCCF-S [50] [ICCV'2021]	52.97	37.02	18.64	43.29	53.81
MCCF-D [50] [ICCV'2021]	64.65	50.45	25.09	51.27	66.51
RSICC [38] [TGRS'2022]	67.92	53.61	25.41	51.96	66.54
SEN [39] [TGRS'2024]	64.12	50.41	23.67	48.19	65.15
PromptCC [76] [TGRS'2023]	70.03	58.41	26.48	55.82	85.44
Change3D [77] [CVPR'2025]	<u>72.25</u>	<u>58.68</u>	<u>27.06</u>	<u>56.04</u>	<u>86.19</u>
SCDistill (Ours)	**72.39**	**59.77**	**29.48**	**59.65**	**95.59**

4.3 Discussions

To further understand the effectiveness of our SCDistill framework, we conduct comprehensive ablation studies and visual analyses focusing on the perturbation simulation, the semantic-invariant self-distillation, few-shot experiments evaluating the generalization ability of the components, the feature-level behavior after self-distillation, and the influence of different teachers in self-distillation.

Effect of Perturbation Simulation. To evaluate the contribution of our diffusion-based perturbation simulation, we remove this component and train the model using only the standard synthetic semantic CD datasets. As shown in Table 4, the absence of perturbation simulation causes a significant drop in performance. Besides, the joint design of our components forms a closed-loop mechanism and achieves larger gains. For example, +2.78 Fscd compared to +0.69 or +1.05 individually, exceeding what would be expected from a simple additive effect. This confirms that constructing a synthetic dataset enriched with non-semantic disturbances provides a more realistic and robust pretraining signal, enabling the model to better distinguish semantic transitions from appearance variations and thus enhancing its generalization to real-world conditions.

Effect of Semantic-invariant Self-distillation. We further evaluate the impact of the semantic-invariant self-distillation strategy by comparing the model using the original VFME without distillation against our self-distilled encoder. As reported in Table 4, removing the self-distillation strategy consistently degrades performance. This demonstrates that the proposed semantic-invariant self-distillation effectively transfers semantic robustness from clean to perturbed representations, guiding the encoder to learn disturbance-resistant semantic features and yielding more reliable and accurate semantic change detection.

Effect of the Main Components via Few-shot Generalization. We further evaluate different ablated settings under a few-shot protocol on the SECOND [62] benchmark to assess generalization from synthetic to real-world data. Specifically, models pretrained with different component combinations are fine-tuned using limited labeled samples. As shown in Table 5, each component contributes to improved few-shot performance, while the full model consistently achieves the best results. This demonstrates that both perturbation simulation and semantic-

Table 4: Effect of the main components.

Setting	Simulate	Distill	Fscd	mIoU	OA	SeK
Vanilla	✗	✗	62.86	72.54	87.53	22.30
+ Simulation	✔	✗	63.55	73.10	87.87	23.39
+ Distillation	✗	✔	63.91	73.13	88.24	23.37
Ours	✔	✔	**65.64**	**74.01**	**88.59**	**25.12**

Table 5: Results of few-shot ablation studies on SECOND [62].

Setting	**1%**				**10%**				**30%**			
	Fscd	mIoU	OA	SeK	Fscd	mIoU	OA	SeK	Fscd	mIoU	OA	SeK
w/o Simulation	51.8	67.6	84.8	11.7	58.9	71.3	87.3	19.1	62.8	72.4	87.5	22.2
w/o Distillation	48.9	67.2	84.2	9.8	59.6	70.0	87.9	18.5	61.8	72.3	87.9	21.4
Ours	**55.6**	**68.7**	**86.0**	**15.0**	**62.5**	**72.4**	**88.0**	**21.6**	**64.4**	**73.3**	**88.4**	**23.5**

invariant self-distillation enhance transferability, enabling strong generalization from synthetic pretraining data to real-world change detection scenarios.

Effect of Self-distillation via Quantitative and Qualitative Analysis. We first provide quantitative evaluation of semantic invariance before and after distillation. Table 6 reports MSE, cosine similarity, and centered kernel alignment (CKA) computed on features extracted from unchanged real-world HRSCD [14] image pairs by encoders before and after distillation. These metrics measure feature discrepancy (MSE), directional consistency (cosine similarity), and representation-level alignment (CKA), respectively. Results show that the distilled encoder achieves lower MSE and higher cosine similarity and CKA, indicating improved robustness to real-world appearance variations.

Furthermore, we visualize the feature differences between bi-temporal inputs as heatmaps, comparing the outputs of the teacher and student models (Fig. 4). The teacher model shows strong responses to illumination, shadows, and other appearance-level variations, leading to spurious activations and weak separation between semantic and non-semantic changes. In contrast, the self-distilled model produces disturbance-resistant and semantically consistent feature differences, effectively distinguishing true semantic transitions from non-semantic fluctuations and providing more robust guidance for change detection.

Effect of Self-distilling by Different Teachers. We further examine the impact of different vision foundation model encoders (VFMEs), including DINOv2 [47] and DINOv3 [55], with and without our semantic-invariant self-distillation, as shown in Table 7. For both teachers, introducing distillation consistently brings substantial performance gains, demonstrating that our semantic-invariant self-distillation effectively improves robustness regardless of the specific teacher model. This indicates that the robustness of the distilled encoder is mainly induced by enforcing semantic consistency across perturbed inputs during distillation, rather than being directly inherited from a robust teacher. Among the teachers, DINOv3 achieves the best overall results, likely due to its

Table 6: Quantitative effect of self-distillation. (%)

Distilling	MSE↓	cosine↑	CKA↑
Before	6.41	84.05	87.08
After	**3.32**	**89.78**	**90.89**

Table 7: Comparison of different teachers w and w/o semantic-invariant self-distillation.

Teacher model	Fscd	mIoU	OA	SeK
DINOv2 [47]	63.09	73.00	87.92	22.80
+ Our Distillation	64.28	73.42	88.41	23.81
DINOv3 [55]	63.55	73.10	87.87	23.39
+ Our Distillation	**65.64**	**74.01**	**88.59**	**25.12**

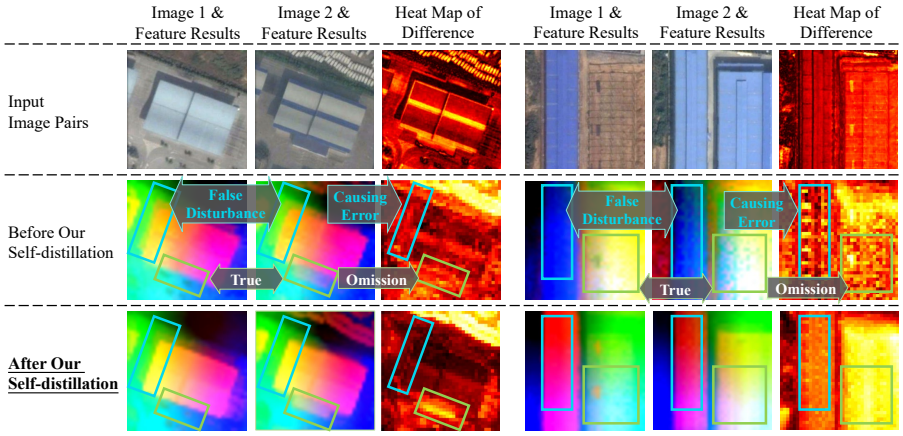

Fig. 4: Comparison of feature representations via PCA [54] before and after semantic-invariant self-distillation on the real-world SECOND dataset [62].

stronger semantic representations on clean natural images, which provide more reliable semantic supervision for our distillation process.

5 Conclusion

In this paper, we introduce **SCDistill**, a framework that enables semantic-robust change detection via semantic-invariant self-distillation. Our approach tackles the limitations of conventional methods, which often mistakenly detect non-semantic variations, such as illumination, shadows, and atmospheric changes, as true semantic changes, leading to degraded robustness in real-world scenarios. Specifically, SCDistill introduces a semantic-invariant self-distillation strategy to transfer semantic robustness from perturbed yet semantic-preserving data, guiding the encoder to learn disturbance-resistant and semantically stable representations. In addition, our method incorporates a diffusion-based perturbation simulation pipeline that synthesizes diverse non-semantic environmental disturbances, allowing the model to explicitly learn to distinguish semantic changes from appearance-level fluctuations. Extensive experiments on multiple benchmarks demonstrate that our SCDistill achieves state-of-the-art performance and strong generalization across various change detection tasks.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62331006), the Fundamental Research Funds for the Central Universities, and the National Natural Science Foundation of China under Grant (625B2026).

References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley interdisciplinary reviews: computational statistics **2**(4), 433–459 (2010)
2. Almahairi, A., Ballas, N., Cooijmans, T., Zheng, Y., Larochelle, H., Courville, A.: Dynamic capacity networks. In: Proceedings of International Conference on Machine Learning. pp. 2549–2558 (2016)
3. Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium. pp. 207–210 (2022)
4. Benidir, Y., Gonthier, N., Mallet, C.: The change you want to detect: Semantic change detection in earth observation with hybrid data generation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 2204–2214 (2025)
5. Bourdis, N., Marraud, D., Sahbi, H.: Constrained optical flow for aerial image change detection. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium. pp. 4176–4179 (2011)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
7. Chang, H., Wang, P., Diao, W., Xu, G., Sun, X.: A triple-branch hybrid attention network with bitemporal feature joint refinement for remote-sensing image semantic change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
8. Chen, H., Li, W., Shi, Z.: Adversarial instance augmentation for building change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2021)
9. Chen, H., Qi, Z., Shi, Z.: Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2021)
10. Chen, H., Shi, Z.: A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote sensing **12** (2020)
11. Chen, H., Song, J., Han, C., Xia, J., Yokoya, N.: Changemamba: Remote sensing change detection with spatiotemporal state space model. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–20 (2024)
12. Chen, T., Lu, Z., Yang, Y., Zhang, Y., Du, B., Plaza, A.: A siamese network based u-net for change detection in high resolution remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 2357–2369 (2022)
13. Cui, F., Jiang, J.: Mtscd-net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images. International Journal of Applied Earth Observation and Geoinformation **118**, 103294 (2023)
14. Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y.: Multitask learning for large-scale semantic change detection. Computer Vision and Image Understanding **187**, 102783 (2019)

15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
16. Ding, L., Guo, H., Liu, S., Mou, L., Zhang, J., Bruzzone, L.: Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)
17. Dong, W., Wang, Y.C., Yang, C., Sun, C., Li, H., Hua, Z., Wu, Z., Chang, X., Bao, L., Qu, S., et al.: Deep learning enhanced in situ atomic imaging of ion migration at crystalline–amorphous interfaces. Nano Letters **24**(45), 14445–14452 (2024)
18. Fang, S., Li, K., Li, Z.: Changer: Feature interaction is what you need for change detection. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–11 (2023)
19. Fang, S., Li, K., Shao, J., Li, Z.: Snunet-cd: A densely connected siamese network for change detection of vhr images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021)
20. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020)
21. Feng, Y., Jiang, J., Xu, H., Zheng, J.: Change detection on remote sensing images using dual-branch multilevel intertemporal network. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)
22. Feng, Y., Xu, H., Jiang, J., Liu, H., Zheng, J.: Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022)
23. Gao, F., Wang, X., Gao, Y., Dong, J., Wang, S.: Sea ice change detection in sar images based on convolutional-wavelet neural networks. IEEE Geoscience and Remote Sensing Letters **16**(8), 1240–1244 (2019)
24. Gu, C., Chen, L., Gu, L., Fu, Y.: Fourier angle alignment for oriented object detection in remote sensing. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 42225–42235 (2026)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
26. Hoxha, G., Chouaf, S., Melgani, F., Smara, Y.: Change captioning: A new paradigm for multitemporal remote sensing image analysis. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)
27. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
28. Ji, S., Wei, S., Lu, M.: Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing **57**, 574–586 (2018)
29. Khan, S.H., He, X., Porikli, F., Bennamoun, M.: Forest change detection in incomplete satellite images with deep neural networks. IEEE Transactions on Geoscience and Remote Sensing **55**(9), 5407–5423 (2017)
30. Lei, T., Geng, X., Ning, H., Lv, Z., Gong, M., Jin, Y., Nandi, A.K.: Ultralightweight spatial-spectral feature cooperation network for change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)
31. Li, H., Wu, Z., Shao, R., Fu, Y.: Statistical characteristic-guided denoising for rapid high-resolution transmission electron microscopy imaging. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 34050–34060 (2026)

32. Li, H., Wu, Z., Shao, R., Zhang, T., Fu, Y.: Noise calibration and spatial-frequency interactive network for stem image enhancement. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 21287–21296 (2025)
33. Li, Z., Wang, X., Fang, S., Zhao, J., Yang, S., Li, W.: A decoder-focused multitask network for semantic change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024)
34. Li, Z., Tang, C., Liu, X., Zhang, W., Dou, J., Wang, L., Zomaya, A.Y.: Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–12 (2023)
35. Liang, Y., Fu, Y.: Relation-guided adversarial learning for data-free knowledge transfer. International Journal of Computer Vision **133**(5), 2868–2885 (2025)
36. Liang, Y., Fu, Y., Hu, Y., Shao, W., Liu, J., Zhang, D.: Flow-anything: Learning real-world optical flow estimation from large-scale single-view images. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(10), 8435–8452 (2025)
37. Liang, Y., Fu, Y., Liu, J., Zhang, D.: Lift3dreamer: Boosting text-driven novel view synthesis via lifted 3d inpainting model from single images. Fundamental Research (2026)
38. Liu, C., Zhao, R., Chen, H., Zou, Z., Shi, Z.: Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–20 (2022)
39. Liu, C., Zhao, R., Chen, J., Qi, Z., Zou, Z., Shi, Z.: A decoupling paradigm with prompt learning for remote sensing image change captioning. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–18 (2023)
40. Liu, M., Chai, Z., Deng, H., Liu, R.: 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 (2022)
41. Liu, W., Lin, Y., Liu, W., Yu, Y., Li, J.: An attention-based multiscale transformer network for remote sensing image change detection. ISPRS Journal of Photogrammetry and Remote Sensing **202**, 599–609 (2023)
42. Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X.: Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geoscience and Remote Sensing Letters **18**(5), 811–815 (2020)
43. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
44. Luo, H., Liu, C., Wu, C., Guo, X.: Urban change detection based on dempster-shafer theory for multitemporal very high-resolution imagery. Remote Sensing **10**(7), 980 (2018)
45. Ma, J., Duan, J., Tang, X., Zhang, X., Jiao, L.: Eatder: Edge-assisted adaptive transformer detector for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2023)
46. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of International Conference on 3D Vision. pp. 565–571 (2016)
47. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

48. Park, D.H., Darrell, T., Rohrbach, A.: Robust change captioning. In: Proceedings of IEEE International Conference on Computer Vision. pp. 4624–4633 (2019)
49. Peng, D., Bruzzone, L., Zhang, Y., Guan, H., He, P.: Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. International Journal of Applied Earth Observation and Geoinformation **103**, 102465 (2021)
50. Qiu, Y., Yamamoto, S., Nakashima, K., Suzuki, R., Iwata, K., Kataoka, H., Satoh, Y.: Describing and localizing multiple changes with transformers. In: Proceedings of IEEE International Conference on Computer Vision. pp. 1971–1980 (2021)
51. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 234–241 (2015)
52. Seo, M., Lee, H., Jeon, Y., Seo, J.: Self-pair: Synthesizing changes from single source for object change detection in remote sensing imagery. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (2023)
53. Shi, S., Zhong, Y., Zhao, J., Lv, P., Liu, Y., Zhang, L.: Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2020)
54. Shlens, J.: A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100 (2014)
55. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
56. Song, J., Chen, H., Yokoya, N.: Syntheworld: A large-scale synthetic dataset for land cover mapping and building change detection. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision. pp. 8287–8296 (2024)
57. Tian, Y., Fu, Y., Zhang, J.: Transformer-based under-sampled single-pixel imaging. Chinese Journal of Electronics **32**(5), 1151–1159 (2023)
58. Vorovencii, I.: A change vector analysis technique for monitoring land cover changes in copsa mica, romania, in the period 1985-2011. Environmental monitoring and assessment **186**, 5951–5968 (2014)
59. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 11534–11542 (2020)
60. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 11379–11388 (2022)
61. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of European Conference on Computer Vision. pp. 3–19 (2018)
62. Yang, K., Xia, G.S., Liu, Z., Du, B., Yang, W., Pelillo, M., Zhang, L.: Asymmetric siamese networks for semantic change detection in aerial images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2021)
63. Yang, Y., Sial, H.A., Baldrich, R., Vanrell, M.: Relighting from a single image: Datasets and deep intrinsic-based architecture. IEEE Transactions on Multimedia (2025)
64. Zang, Q., Yang, J., Wang, S., Zhao, D., Yi, W., Zhong, Z.: Changediff: A multitemporal change detection data generator with flexible text prompts via diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9763–9771 (2025)

65. Zhang, R., Zhang, H., Ning, X., Huang, X., Wang, J., Cui, W.: Global-aware siamese network for change detection on remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing **199**, 61–72 (2023)
66. Zhang, T., Fu, Y., Zhang, J., Yan, C.: Deep guided attention network for joint denoising and demosaicing in real image. Chinese Journal of Electronics **33**(1), 303–312 (2024)
67. Zhang, Y., Chen, S., Tian, Y., Gao, Y., Jiang, J., Fu, Y.: Supervise-assisted multi-modality fusion diffusion model for pet restoration. Tsinghua Science and Technology (2026)
68. Zhang, Y., Lai, Z., Zhang, T., Fu, Y., Zhou, C.: Unaligned rgb guided hyperspectral image super-resolution with spatial-spectral concordance. International Journal of Computer Vision **133**(9), 6590–6610 (2025)
69. Zhang, Y., Zhang, T., Nie, J., Fu, Y.: Enhancing unregistered hyperspectral image super-resolution via unmixing-based abundance fusion learning. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 41573–41583 (2026)
70. Zhang, Y., Zhang, T., Nie, J., Fu, Y.: Real noise decoupling for hyperspectral image denoising. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12925–12933 (2026)
71. Zheng, Z., Ermon, S., Kim, D., Zhang, L., Zhong, Y.: Changen2: Multi-temporal remote sensing generative change foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(2), 725–741 (2024)
72. Zheng, Z., Tian, S., Ma, A., Zhang, L., Zhong, Y.: Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In: Proceedings of IEEE International Conference on Computer Vision. pp. 21818–21827 (2023)
73. Zheng, Z., Tian, S., Ma, A., Zhang, L., Zhong, Y.: Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In: Proceedings of IEEE International Conference on Computer Vision. pp. 21818–21827 (2023)
74. Zheng, Z., Zhong, Y., Tian, S., Ma, A., Zhang, L.: Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. ISPRS Journal of Photogrammetry and Remote Sensing **183**, 228–239 (2022)
75. Zheng, Z., Zhong, Y., Zhang, L., Ermon, S.: Segment any change. In: Proceedings of Advances in Neural Information Processing Systems (2024)
76. Zhou, Q., Gao, J., Yuan, Y., Wang, Q.: Single-stream extractor network with contrastive pre-training for remote-sensing change captioning. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
77. Zhu, D., Huang, X., Huang, H., Zhou, H., Shao, Z.: Change3d: Revisiting change detection and captioning from a video modeling perspective. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 24011–24022 (2025)
78. Zhu, T., Li, H., Fu, Y.: Trim-sod: A multi-modal, multi-task, and multi-scale spacecraft optical dataset. Space: Science Technology **5**, 0299 (2025)