

A semantic prior-enhanced semi-supervised method for on-board change detection

Kexin Zhu¹, Zizhen Zhang², Jun Pan^{1,3,4}

¹State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China, ²School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China, ³Oriental Space Port Research Institute, Yantai 265100, China, ⁴Hubei LuoJia Laboratory, 430072, Wuhan, China

Corresponding author: Jun Pan, panjun1215@whu.edu.cn

Change Detection (CD) plays a crucial role in Earth observation applications such as environmental assessment, land use analysis and disaster monitoring. As space information systems increasingly demand on-board data processing capabilities, CD models are required to deliver reliable performance under conditions where labelled samples are scarce and data distributions are unknown. Conventional supervised approaches, however, rely heavily on large-scale annotated datasets and often generalize poorly to unseen on-board data, resulting in noticeable performance degradation. To address these limitations, this paper proposes a semantic prior-enhanced semi-supervised method for on-board CD. The proposed approach incorporates general semantic priors derived from vision foundation models, thereby alleviating the limited semantic representation capacity caused by insufficient annotations and improving the detection of complex real-world changes. To balance the integration of rich semantic priors with the strict computational constraints of on-board scenarios, a semantic prior-enhanced encoder based on DINOv3 fine-tuning is proposed, alongside a lightweight frequency-separated feature adaptation mechanism. This design preserves transferable visual knowledge while substantially reducing fine-tuning overhead, enabling efficient adaptation to the domain discrepancies of remote sensing imagery and yielding more discriminative feature representations. In addition, a complementary cross-temporal feature perturbation strategy is proposed to better exploit latent information within unlabelled data and encourage the learning of robust change features. Experiments conducted on three public datasets and real-world scenarios demonstrate that the proposed method consistently outperforms existing state-of-the-art approaches under limited annotation settings. Using only 5% labelled samples, the proposed method improves the F1-score by 2.44%, 2.70% and 1.48%, respectively, confirming its effectiveness and generalization capability in on-board CD tasks.

Keywords – Change detection, deep learning, remote sensing, semi-supervised learning, vision foundation models

1. INTRODUCTION

Change Detection (CD) is one of the fundamental tasks in remote sensing image understanding. Its core objective is to identify changes in land cover and surface conditions by analysing and comparing multi-temporal remote sensing images acquired over the same geographic region [1]. With advances in Earth observation technology, CD has become essential for Earth system monitoring and is widely applied to urban expansion analysis, disaster assessment and ecological monitoring [2–5]. In recent years, within modern Earth observation frameworks, CD has gradually shifted from conventional offline analysis toward the more challenging paradigm of on-board real-time processing. On-board systems are typically required to perform data processing under resource-constrained conditions while coping with unknown and continuously evolving data distributions, which imposes stricter demands on both model efficiency and generalization capability. The rapid development of high-resolution satellites and multi-platform remote sensing systems has led to massive volumes of remote sensing data, making efficient and reliable on-board CD an important research direction [6–8].

In recent years, deep learning-based methods have substantially advanced the field of CD. Among them, Convolutional Neural Networks (CNNs), owing to their local receptive fields and parameter-sharing mechanism, have served as the core backbone of early mainstream approaches. Typical CNN-based methods adopt a Siamese architecture to extract bi-temporal features separately and model changes through feature differences [9–11]. As research has progressed, the transformer architecture has been introduced to strengthen global dependency modelling

and has been combined with CNNs to form hybrid architectures that further improve performance [12-14]. In parallel, semi-supervised CD has gradually emerged as an active research focus. Semi-supervised CD aims to exploit a small amount of labelled data together with large-scale unlabelled data, thereby alleviating the high cost of pixel-level annotation [15]. Existing semi-supervised methods can generally be divided into three categories: pseudo-label learning [16], consistency regularization [17] and teacher-student frameworks [18]. Pseudo-label methods generate supervisory signals from model predictions to expand the training set, but they are prone to error accumulation caused by incorrect predictions. Consistency regularization enhances robustness by enforcing prediction consistency under different perturbations and has been widely adopted in remote sensing CD. The teacher-student framework, in turn, maintains a stable teacher model via Exponential Moving Average (EMA), providing more reliable supervisory signals for the student model.

Although existing methods have achieved notable progress, they still face far more severe challenges in practical on-board remote sensing applications. In real on-board environments, the data distribution is typically unknown and differs significantly from that of the training data, resulting in severe domain shift [19]. Traditional fully supervised methods rely on large-scale annotated data and perform well on in-distribution datasets, yet their generalization ability is markedly limited in such scenarios, leading to substantial performance degradation. In contrast, semi-supervised learning exploits unlabelled target-domain data and consistency constraints to improve robustness under distribution shifts. This enhances generalization to unseen distributions and offers a promising direction for on-board CD [20]. Nevertheless, current semi-supervised CD methods still suffer from limitations in pseudo-label quality control and semantic information exploitation. As a result, models remain susceptible to noisy labels and struggle to learn stable and highly discriminative change features [21-23].

In recent years, vision foundation models have acquired powerful general purpose semantic representations through self-supervised learning on large-scale unlabelled data, demonstrating strong generalization across a wide range of visual tasks [24, 25]. Representative models such as CLIP, SAM and the DINO series have shown remarkable performance in zero-shot learning and cross-task transfer [26-28]. Among these, the DINO family (e.g. DINOv2 and DINOv3) exhibits exceptional feature extraction capability in few-shot representation learning [29, 30]. Unlike conventional models trained from scratch, foundation models capture class-agnostic structural information and long-range dependencies from large-scale data, yielding highly discriminative semantic representations even under limited supervision. More recently, researchers have begun to introduce such vision foundation models into the remote sensing field, aiming to exploit their rich semantic priors to enhance model understanding of complex land cover scenes [31, 32]. However, owing to substantial differences between natural images and remote sensing imagery in imaging perspective, spatial structure and texture distribution,

directly transferring vision foundation models to on-board CD still suffers from notable domain shift, which hinders the effective use of their semantic prior knowledge.

To address these challenges, this paper proposes a Semantic Prior-Enhanced semi-supervised method for on-board Change Detection (SPE-CD). By incorporating semantic priors from vision foundation models, the proposed method strengthens the model's ability to represent change-related semantics under unknown data distributions, thereby improving generalization and detection accuracy in on-board scenarios. Specifically, given the substantial domain gap between natural images and remote sensing imagery, we design a Semantic Prior-Enhanced Encoder (SPE-Encoder) built upon a frequency-separated feature adaptation mechanism. This encoder adopts the pre-trained vision foundation model DINOv3 as its backbone and incorporates a lightweight decoupled feature adaptation module. While preserving general semantic knowledge, the module performs frequency band-wise calibration via spatial frequency decoupling, yielding task-aligned semantic representations with improved pseudo-label reliability. Building on this, and to accommodate the bi-temporal nature of CD, we further propose a Complementary Cross-temporal Feature Perturbation (CCFP) strategy. This strategy constructs multi-view complementary perturbations at the feature level and enforces cross-temporal consistency constraints, encouraging the model to learn stable and highly discriminative change representations. The main contributions of our work are as follows.

1. This paper proposes the SPE-CD, which leverages the semantic priors of DINOv3 to improve generalization under unknown data distributions.
2. This paper designs the SPE-encoder with frequency-separated feature adaptation to reduce the domain gap between natural and remote sensing imagery through spatial frequency decoupling and frequency band-wise calibration, thereby improving pseudo-label reliability.
3. This paper proposes the CCFP strategy that constructs complementary multi-view perturbations and enforces cross-temporal consistency constraints to learn stable and discriminative change representations.

2. METHODOLOGY

2.1 Overall structure of SPE-CD

To address insufficient exploitation of unlabelled semantic information and the domain shift of vision foundation models in remote sensing scenarios, this paper proposes a semantic prior-enhanced semi-supervised on-board change detection method. Built upon a teacher-student framework, the proposed method integrates semantic prior modelling with structured feature perturbation to improve semantic discrimination and robustness under limited supervision. Specifically, an SPE-Encoder based on spatial-frequency decoupling is designed to adapt intermediate features of the vision foundation model to remote sensing imagery while preserving general semantic priors. In addition, a CCFP strategy is proposed to construct complementary multi-view

representations and strengthen robust modelling of change relationships. A change-guided consistency learning strategy is further incorporated to improve pseudo-label quality and enhance the utilization of unlabelled data.

Fig. 1 illustrates the overall framework of the proposed SPE-CD method. The SPE-Encoder extracts discriminative semantic features, while the CCFP strategy enhances

consistency learning through multi-view perturbations in the student branch. The framework is trained using both supervised learning on labelled samples and semi-supervised learning on unlabelled samples. During semi-supervised training, the teacher model generates pseudo-labels for unlabelled data and the student model is optimized under multiple perturbations to progressively improve robustness and generalization.

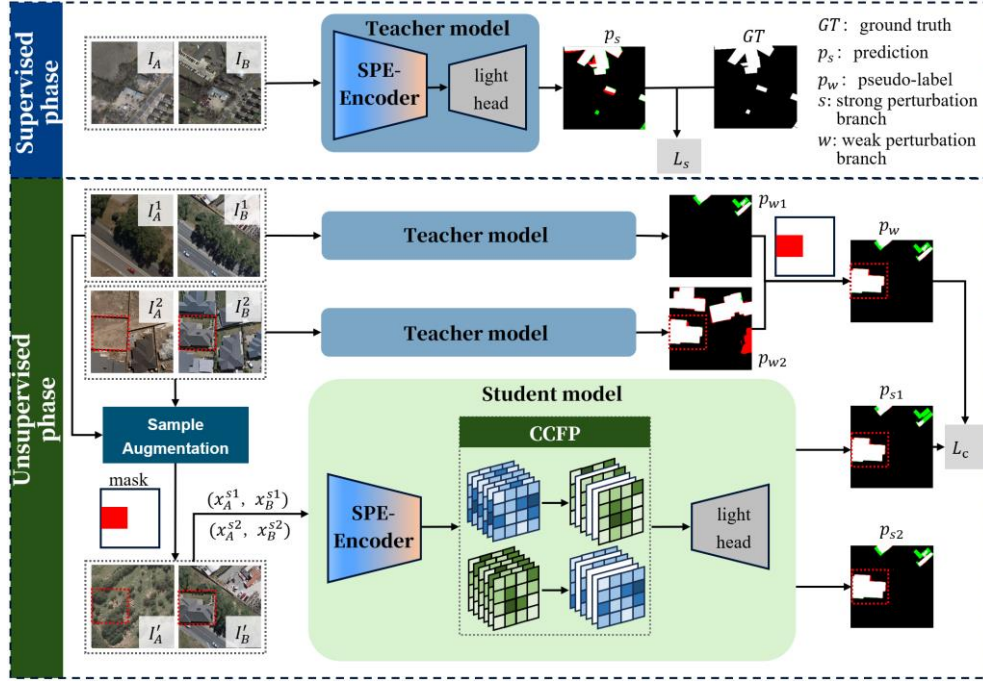


Figure 1 – The overall framework of the SPE-CD.

2.2 Semi-supervised training pipeline

In the supervised learning stage, labelled sample pairs (I_A, I_B) are processed by the SPE-Encoder and the prediction head to generate change maps supervised by the cross-entropy loss:

$$L_s = L_{ce}(p_s, GT) \quad (1)$$

where p_s denotes the predicted change map of the student model and GT represents the ground-truth label. This stage mainly provides stable initialization for subsequent semi-supervised learning.

In the semi-supervised stage, a teacher–student framework is adopted, where the teacher model generates pseudo-labels and the student model performs consistency learning under multiple perturbations. Given two unlabelled sample pairs (I_A^1, I_B^1) and (I_A^2, I_B^2) , the teacher first produces prediction maps (p_{w1}, p_{w2}) , which are used to construct change-guided mixed samples through a CutMix-based augmentation strategy. Specifically, a change saliency map is derived from the teacher predictions to guide mask generation, increasing the probability that mixed regions contain changed objects while maintaining low computational overhead. The mixed samples are formulated as:

$$I'_A = I_{A1} \odot M + I_{A2} \odot (1 - M) \quad (2)$$

$$I'_B = I_{B1} \odot M + I_{B2} \odot (1 - M) \quad (3)$$

where $\odot$ denotes the element-wise multiplication operation and M represents the change-guided binary mask. The corresponding pseudo-labels are fused using the same masking rule to obtain the mixed pseudo-label p_w .

To improve robustness under perturbations, the mixed samples are processed by two independent strong augmentations and fed into the student model. After semantic feature extraction, the proposed CCFP strategy is applied to construct complementary channel perturbations and random temporal feature swapping, producing predictions p_{s1} and p_{s2} . A consistency loss is then imposed between the student predictions and the mixed pseudo-label:

$$L_c = \frac{1}{2} (\|p_{s1} - p_w\|_2^2 + \|p_{s2} - p_w\|_2^2) \quad (4)$$

The overall training objective is defined as:

$$L = \frac{1}{2} (L_s + L_c) \quad (5)$$

The teacher model parameters are updated using EMA:

$$\theta_t = \alpha \theta_{t-1} + (1 - \alpha) \theta_s \quad (6)$$

where θ_t and θ_s denote the teacher and student parameters, respectively, and α is the smoothing coefficient.

2.3 Semantic Prior-Enhanced Encoder (SPE-Encoder)

To fully inherit the general semantic priors learned by vision foundation models from large-scale natural images, while alleviating the domain gap between natural and remote sensing imagery, we propose an SPE-Encoder built upon a spatial-frequency decoupling mechanism. The encoder adopts a frozen pre-trained DINOv3 backbone as its base and introduces lightweight feature adapters after each transformer block. Without updating backbone parameters, only the adapters are optimized, enabling parameter-efficient adaptation that preserves pre-trained semantic knowledge while progressively aligning intermediate features with remote sensing data distributions. The overall architecture is illustrated in Fig. 2.

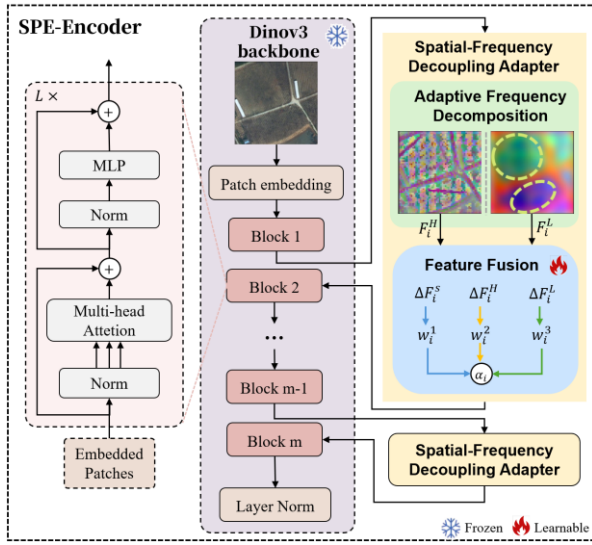


Figure 2 – Schematic diagram of the SPE-Encoder based on spatial frequency decoupling.

Let the input remote sensing image be denoted as $X \in \mathbb{R}^{B \times C \times H \times W}$. After patch embedding and transformer encoding, the output feature of the i -th layer is denoted as $F_i \in \mathbb{R}^{(hw) \times c}$, where hw is the token length and c is the channel dimension. For frequency-domain processing, F_i is reshaped into $F_i \in \mathbb{R}^{C \times H \times W}$, where $C = c$ and $hw = H \times W$. To enhance parameter-efficient adaptation, we introduce a spatial-frequency decoupled mechanism, as illustrated in Fig. 3. In the spatial branch, a low-rank adapter is applied to model spatial transformations:

$$\Delta F_i^s = \mathcal{A}_s^{(i)}(F_i) \quad (7)$$

where the spatial adapter $\mathcal{A}_s^{(i)}(\cdot)$ is defined as:

$$\mathcal{A}_s^{(i)}(X) = W_s^{up} \sigma(W_s^{down} X) \quad (8)$$

where $W_s^{down} \in \mathbb{R}^{r \times C}$ and $W_s^{up} \in \mathbb{R}^{C \times r}$ denote the down-projection and up-projection matrices, respectively, $r \ll C$ is the bottleneck dimension, and $\sigma(\cdot)$ represents the non-linear activation function (e.g. ReLU). Through this low-rank factorization, the semantic features in the spatial domain can be effectively enhanced while keeping the parameter count low.

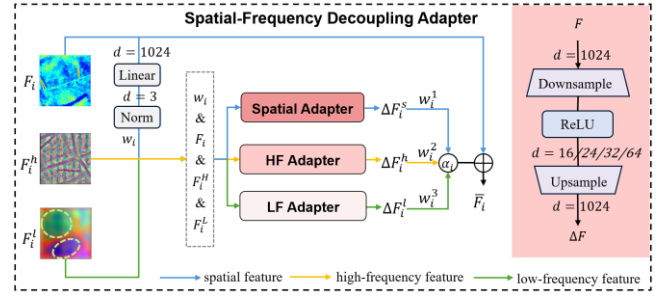


Figure 3 – Schematic of the spatial-frequency decoupling adapter.

To further exploit the structural information of visual features in the frequency domain, we design a frequency-decoupled feature adaptation mechanism. Visual features in different frequency subspaces tend to correspond to different semantic levels. Low frequency components mainly encode coarse-grained information such as the overall structure and spatial layout of land cover features, whereas high frequency components focus on capturing fine-grained information such as edge contours and detailed textures. Accordingly, the proposed method adaptively decomposes features in the frequency domain, as illustrated in Fig. 4. Specifically, the feature F_i is first reshaped into a three-dimensional tensor: $X_i \in \mathbb{R}^{C \times H \times W}$, after which a two-dimensional discrete Fourier transform is independently applied along the spatial dimensions for each channel to obtain the spectral representation:

$$\hat{X}_i = \mathcal{F}(X_i) \quad (9)$$

To separate the low and high frequency information, we define a frequency mask function $M_{\rho_i} \in \{0,1\}^{H \times W}$, where $\rho_i \in (0,1)$ denotes the cu-toff ratio. Unlike conventional approaches that partition high and low frequencies using a fixed threshold, we design ρ_i as a learnable parameter constrained by a sigmoid function:

$$\rho_i = \text{Sigmoid}(\theta_i) \quad (10)$$

where θ_i is a trainable variable. In practice, we implement the mask via a soft thresholding operation as follows:

$$M_{\rho_i} = \text{Sigmoid}((\rho_i - d)/\tau) \quad (11)$$

where d is the normalized frequency radius and $\tau = 2.0$ controls the transition sharpness. This formulation is fully differentiable, enabling standard backpropagation for θ_i . Accordingly, the low and high frequency spectra are expressed as:

$$\hat{X}_i^l = M_{\rho_i} \odot \hat{X}_i \quad (12)$$

$$\hat{X}_i^h = (1 - M_{\rho_i}) \odot \hat{X}_i \quad (13)$$

where $\odot$ denotes element-wise multiplication. They are then mapped back to the feature space via the inverse Fourier transform:

$$F_i^l = \mathcal{F}^{-1}(\hat{X}_i^l) \quad (14)$$

$$F_i^h = \mathcal{F}^{-1}(\hat{X}_i^h) \quad (15)$$

Once the low and high frequency features are obtained, a low frequency adapter $\mathcal{A}_l^{(i)}$ and a high frequency adapter $\mathcal{A}_h^{(i)}$ are constructed. Their structures are identical to that of the spatial adapter, but their parameters are independent:

$$\Delta F_i^l = \mathcal{A}_l^{(i)}(F_i^l) \quad (16)$$

$$\Delta F_i^h = \mathcal{A}_h^{(i)}(F_i^h) \quad (17)$$

Through the design of the frequency-decoupled adapters, the model can perform targeted calibration at both the coarse-grained low frequency level and the fine-grained high frequency level.

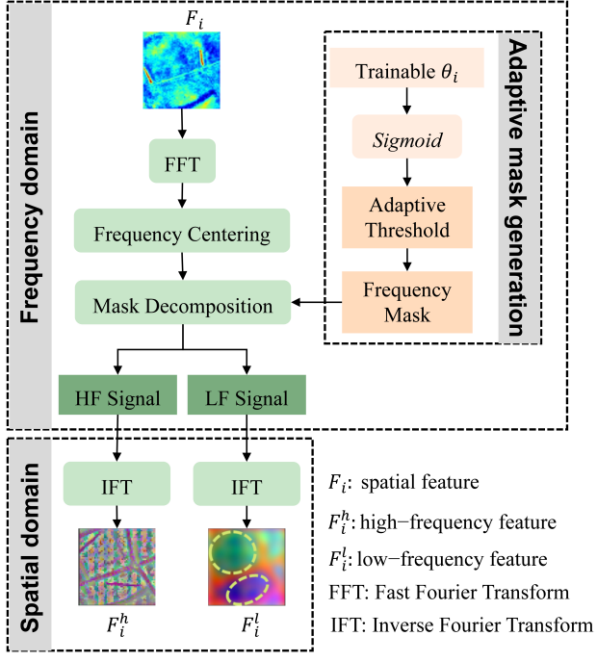


Figure 4 – Schematic of the adaptive frequency decomposition procedure.

To fuse multi-branch information, a dynamic routing mechanism is introduced. The fusion weights are computed as:

$$\omega_i = \text{Softmax}(\mathcal{R}(\text{GAP}(F_i))) \quad (18)$$

where $\text{GAP}(\cdot)$ denotes global average pooling and $\mathcal{R}(\cdot)$ is a linear projection producing three weights corresponding to spatial, low frequency and high frequency branches.

The final feature refinement is given by:

$$\Delta F_i = \beta_i(\omega_i^s \Delta F_i^s + \omega_i^l \Delta F_i^l + \omega_i^h \Delta F_i^h) \quad (19)$$

where β_i is a learnable scaling factor initialized to a small value to stabilize early training. The fused feature is obtained via residual connection:

$$\tilde{F}_i = F_i + \Delta F_i \quad (20)$$

The fused feature $\tilde{F}_i$ is then passed to the next block for further feature extraction.

We provide a parameter cost analysis to quantitatively demonstrate the efficiency of the design. The SPE-Encoder adopts a frozen DINOv3-base backbone with 86.56M

parameters. The trainable components include the spatial, low frequency, and high frequency adapters, the learnable cut-off ratios θ_i , the scaling factors β_i and the dynamic routing modules, which together amount to 7.19M parameters. Thus, the full model contains 93.76M parameters in total, with a freeze ratio of 92.3% (i.e. only 7.7% of the parameters are trainable). Each adapter itself, based on low-rank factorization with $r = 8$, introduces merely thousands of parameters per transformer layer, which is negligible compared to the 86.56M backbone. This confirms that our frequency-decoupled adaptation mechanism achieves effective domain adaptation with minimal parameter overhead.

Unlike uniform fine-tuning of DINO, CLIP, or other vision foundation models, our approach explicitly accounts for the frequency-wise characteristics of the domain gap between natural and remote sensing imagery. In remote sensing, low frequency components primarily encode large-scale land cover structures and overall scene layouts, high frequency components capture fine structural details such as building edges, road boundaries and texture patterns, which differ markedly between natural objects and anthropogenic structures. By decoupling adaptation along spatial and frequency dimensions with independent lightweight adapters and learnable cut-off ratios per layer, our design enables targeted calibration at different semantic granularities, effectively mitigating cross-domain discrepancies while preserving the rich general semantic priors encoded in the frozen backbone.

2.4 Complementary Cross-Temporal Feature Perturbation (CCFP) strategy

To further enhance feature learning on unlabelled samples, we propose a CCFP strategy within a consistency learning framework. Conventional random dropout or single-branch perturbations typically operate independently within each branch and lack structural design, failing to fully exploit the inherent correlations between bi-temporal features in CD. To address this, we introduce a structured feature-level perturbation mechanism that is both cross-temporal and complementary.

Specifically, two strong-perturbation branches are first constructed for each sample, and complementary channel masks are generated so that each branch attends to different feature subspaces, forming multi-view complementary representations. A binary mask is then introduced to randomly swap bi-temporal features, imposing a cross-temporal perturbation that encourages stable change discrimination under varying temporal configurations. This design provides diverse and complementary views in feature space, improving the robustness and generalization of change representation learning in semi-supervised settings, as illustrated in Fig. 5.

For the input bi-temporal images I_A and I_B , two different image-level strong data perturbations (e.g. random flipping and colour jittering) are applied. The SPE-Encoder then extracts the semantically enriched bi-temporal features $F_A^{s1}, F_B^{s1} \in \mathbb{R}^{N \times c \times h \times w}$, $F_A^{s2}, F_B^{s2} \in \mathbb{R}^{N \times c \times h \times w}$ from the two

perturbation branches s_1 and s_2 , respectively. Here, N denotes the batch size, c the number of channels, h, w the spatial dimensions of the features, and the superscripts s_1 and s_2 indicate the two different strong-perturbation branches.

To enable the two perturbation branches to learn change information from complementary feature subspaces, we introduce a complementary feature perturbation mechanism along the feature channel dimension. The core idea is to construct two complementary feature views through channel-wise random perturbation while preserving the structural consistency of the bi-temporal features. This encourages the model to learn change information from different feature subspaces, thereby enhancing the diversity and robustness of feature representation. Specifically, a channel mask is first randomly generated within each batch. For sample $i \in \{1, \dots, N\}$ and channel $j \in \{1, \dots, C\}$, an independent Bernoulli random variable is introduced:

$$U_{i,j} \sim \text{Bernoulli}(0.5), i.i.d. \quad (21)$$

and the channel mask for branch s_1 is then constructed as:

$$m_{i,j}^{s_1} = 2U_{i,j} \quad (22)$$

where $m_{i,j}^{s_1} = 2$ indicates that the channel is preserved, while $m_{i,j}^{s_1} = 0$ indicates that it is suppressed. To preserve the expected value of the features before and after perturbation, the Bernoulli random variable is multiplied by a factor of 2. The Bernoulli distribution is adopted here because the perturbation operation requires a binary decision for each channel. Unlike continuous distributions that would necessitate additional discretization, the Bernoulli distribution directly yields a binary mask in a natural and parameter-free manner. The scaling factor of 2 further ensures that the expected value of the masked features remains unchanged, thus preserving the overall feature magnitude and preventing training instability. The complementary mask for the branch s_2 is then constructed as:

$$m_{i,j}^{s_2} = 2 - m_{i,j}^{s_1} \quad (23)$$

so that:

$$m_{i,j}^{s_1} + m_{i,j}^{s_2} = c \quad (24)$$

where $c = 2$. As can be seen, the two perturbation branches form a strictly complementary relationship along the channel dimension: a channel preserved in one branch is suppressed in the other, and a channel suppressed in one branch is activated in the other.

On this basis, to prevent overly strong feature perturbations from destabilizing model training, we further introduce a random retention strategy. A subset $\mathcal{K}$ of size $\lfloor N/2 \rfloor$ is uniformly sampled from the batch index set $\{1, \dots, N\}$. For the selected samples $i \in \mathcal{K}$, no channel perturbation is applied; that is:

$$m_{i,j}^{s_1} = 1, m_{i,j}^{s_2} = 1 \quad (25)$$

In this case, the sample features are propagated through both branches in their complete form, providing the model with a

stable supervisory signal and partially alleviating over-regularization. This procedure yields the channel masks M^{s_1} and M^{s_2} for the two branches.

Unlike conventional perturbation methods, the same channel masks are shared across both temporal phases, preserving bi-temporal alignment in the feature space. The resulting perturbed features are:

$$\widehat{F}_A^{s_1} = F_A^{s_1} \odot M^{s_1}, \widehat{F}_B^{s_1} = F_B^{s_1} \odot M^{s_1} \quad (26)$$

$$\widehat{F}_A^{s_2} = F_A^{s_2} \odot M^{s_2}, \widehat{F}_B^{s_2} = F_B^{s_2} \odot M^{s_2} \quad (27)$$

where $M^{s_k} \in \mathbb{R}^{N \times C \times 1 \times 1}$ is broadcast over spatial dimensions, and $\odot$ denotes element-wise multiplication. This enables learning from complementary subspaces while preserving temporal consistency.

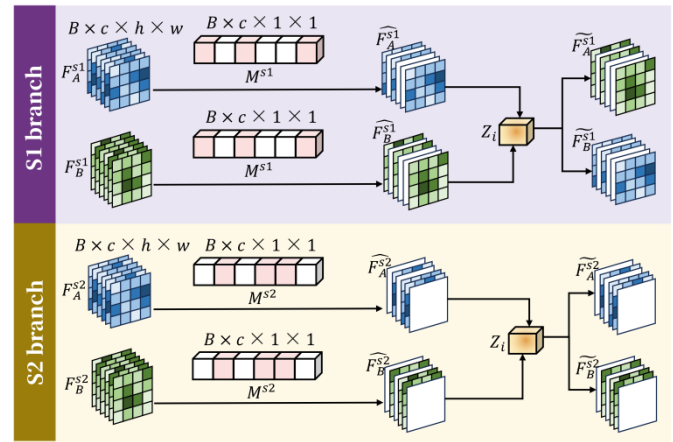


Figure 5 – Schematic of the CCFP strategy.

After obtaining the feature representations from the two complementary feature-level perturbation branches, we further introduce cross-temporal perturbation at the temporal level. The CD task inherently exhibits temporal consistency: regardless of the order in which the two temporal phases are combined as input, the predicted change regions should remain the same. Therefore, additional perturbation views can be constructed by randomly shuffling the temporal order, which strengthens the model's robustness to temporal permutations. Specifically, within any perturbation branch s_k ($k = 1, 2$), a random temporal swap operation is applied to the corresponding bi-temporal features $\widehat{F}_A^{s_k}$ and $\widehat{F}_B^{s_k}$. For sample i in the batch, a binary random variable is introduced:

$$Z_i \sim \text{Bernoulli}(0.5) \quad (28)$$

where $Z_i = 1$ indicates that the cross-temporal swap is performed, and $Z_i = 0$ indicates that the original temporal order is retained. According to this random variable, the two temporal features are transformed as follows:

$$\tilde{F}_A^{s_k} = \begin{cases} \widehat{F}_A^{s_k}, & Z_i = 0 \\ \widehat{F}_B^{s_k}, & Z_i = 1 \end{cases} \quad (29)$$

$$\tilde{F}_B^{s_k} = \begin{cases} \widehat{F}_B^{s_k}, & Z_i = 0 \\ \widehat{F}_A^{s_k}, & Z_i = 1 \end{cases} \quad (30)$$

where $\tilde{F}_A^{sk}$ and $\tilde{F}_B^{sk}$ denote the feature representations after cross-temporal perturbation. Through this operation, the model can observe different temporal arrangements during training and thereby learn change-discriminative features that are insensitive to temporal order. It is worth noting that this perturbation does not alter the true change semantics, since the goal of CD depends solely on the difference between two temporal phases rather than on their input order. By introducing random temporal swapping at the feature level, the perturbation perspectives are further enriched and act synergistically with the mentioned complementary feature perturbation, thereby enhancing the model's generalization and robustness in semi-supervised settings.

3. EXPERIMENTS AND ANALYSIS

3.1 Experimental setup

This section presents the experiments and analyses of the proposed SPE-CD, evaluating its performance from both quantitative and qualitative perspectives through comparative experiments, ablation studies and experiments in real-world scenarios. We first introduce the three experimental datasets and the corresponding parameter settings. In the comparative experiments, the proposed method is then benchmarked against eight state-of-the-art CD methods. Subsequently, ablation analyses are conducted on the core designs of the proposed method, including the encoder design, the perturbation strategy and the fine-tuning strategy. Finally, experiments in real-world scenarios are carried out to further verify the effectiveness and generalization capability of the proposed method.

3.1.1 Evaluation metrics

To quantitatively assess the performance of the proposed method, four widely used evaluation metrics are adopted: Precision (Pre), Recall (Rec), F1-score (F1) and Intersection over Union (IoU). The mathematical formulations of these metrics are specified below:

$$Pre = \frac{TP}{(TP + FP)} \quad (31)$$

$$Rec = \frac{TP}{(TP + FN)} \quad (32)$$

$$F1 = \frac{2Pre \times Rec}{Pre + Rec} \quad (33)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (34)$$

where True Positive (TP) and True Negative (TN) denote the successfully classified changed and unchanged pixels, respectively, while False Positive (FP) represents unchanged regions falsely marked as changes and False Negative (FN) corresponds to actual changes overlooked by the model. Specifically, Pre indicates the capacity of the network to minimize false alarms, meaning elevated Pre values correspond to fewer false detections. Conversely, Rec characterizes the efficacy in avoiding missed detections, where higher Rec denotes a lower omission rate. F1 serves as the harmonic mean that balances Pre and Rec, offering a comprehensive reflection of the model's overall performance. Finally, IoU quantifies the spatial overlap between the predicted segmentations and the ground truth, with superior IoU reflecting enhanced geometric accuracy.

3.1.2 Implementation details

All experiments in this paper were conducted on an NVIDIA Tesla P100 GPU (16 GB) and implemented using the PyTorch deep learning framework. All network parameters were initialized from a zero-mean Gaussian distribution with a standard deviation of 0.02. The network was trained with the SGD optimizer, using a momentum of 0.9 and a weight decay of 0.0005. The total number of training epochs was set to 80, and the batch size was set to 8 for both the labelled and unlabelled training sets. The initial learning rate was set to 0.01 and then linearly decayed to 0. For the data augmentation perturbations involved in our method and in the compared semi-supervised methods (e.g. UniMatch), the strong augmentation strategy proposed by Cubuk et al. was adopted to perturb the unlabelled remote sensing image pairs. The available operations include identity transformation, contrast adjustment, auto-contrast, histogram equalization, brightness adjustment, colour adjustment, posterization, sharpness enhancement and colour inversion [33]. To prevent overly strong perturbations from destroying the semantic information of the images, only two operations are randomly selected from the above set and combined in each application.

3.2 Datasets

To validate the effectiveness of SPE-CD, we select three representative publicly available remote sensing CD datasets: LEVIR-CD [34], WHU-CD [35] and SYSU-CD [36]. These three datasets are chosen because they are widely used in the remote sensing CD community and are highly representative, providing a unified and comparable experimental benchmark for evaluating model performance. The basic information of these three datasets is summarized in Table 1. A detailed description of each dataset is provided below.

Table 1 – Comparison of information from the three used CD datasets.

Datasets	Image-size	Space-resolution (m/pixel)	Image- pairs	Train:Val:Test	Image-type	Changing-objects
LEVIR-CD	1024×1024	0.5	637	7:1:2	Satellite image	building
WHU-CD	32 207×15 354	0.2	1	8:1:1	Aerial image	building
SYSU-CD	256×256	0.5	20 000	6:2:2	Aerial image	Multiple types

The LEVIR-CD dataset consists of 637 pairs of bi-temporal high-resolution remote sensing images with a spatial resolution of 0.5 m/pixel and an image size of 1024×1024 pixels, primarily collected from the Google Earth platform. It covers multiple urban regions across the state of Texas, USA, spanning 20 different areas with time intervals ranging from approximately 5 to 14 years, and therefore contains pronounced land-surface change information. The change types are dominated by building changes, including residential buildings, storage facilities and auxiliary structures, amounting to 31 333 individual building change instances. Considering computational constraints, the original images are typically cropped into non-overlapping 256×256 sub-patches. These sub-patches are then split into training, validation and test sets according to a fixed ratio (e.g. 7:1:2), yielding 7,120/1,024/2,048 pairs, respectively, to meet the requirements of model training and evaluation.

The WHU-CD dataset, released by Wuhan University, focuses primarily on building CD. It is built around the urban reconstruction following the 2011 Christchurch earthquake, with remote sensing images acquired in 2012 and 2016 forming the bi-temporal pairs. The spatial resolution is approximately 0.2 m/pixel, and the image quality is relatively high. The annotations indicate that the two

temporal phases contain 12 796 and 16 077 building instances, respectively, effectively reflecting both the addition and the modification of buildings. Owing to the large size of the original images, they are cropped into non-overlapping 256×256 patches and randomly partitioned into training, validation and test sets at a ratio of 8:1:1, corresponding to 5 947/743/744 pairs for model training, validation and testing.

The SYSU-CD dataset, released by Sun Yat-sen University, covers the Hong Kong region with images acquired between 2007 and 2014. It comprises 20 000 pairs of bi-temporal remote sensing image patches, each of size 256×256 pixels with a spatial resolution of 0.5 m/pixel. The dataset is pre-partitioned into training, validation and test sets at a ratio of 6:2:2, corresponding to 12 000, 4 000 and 4 000 sample pairs, respectively. The dataset encompasses a variety of complex change types, such as road expansion, newly constructed buildings, vegetation changes and suburban expansion, and it also includes fine-grained scenes such as foundation changes during early-stage construction. This diverse set of change patterns makes the dataset suitable not only for evaluating CD models but also for supporting studies on urban development and land use evolution. Sample image pairs from these datasets are shown in Fig. 6.

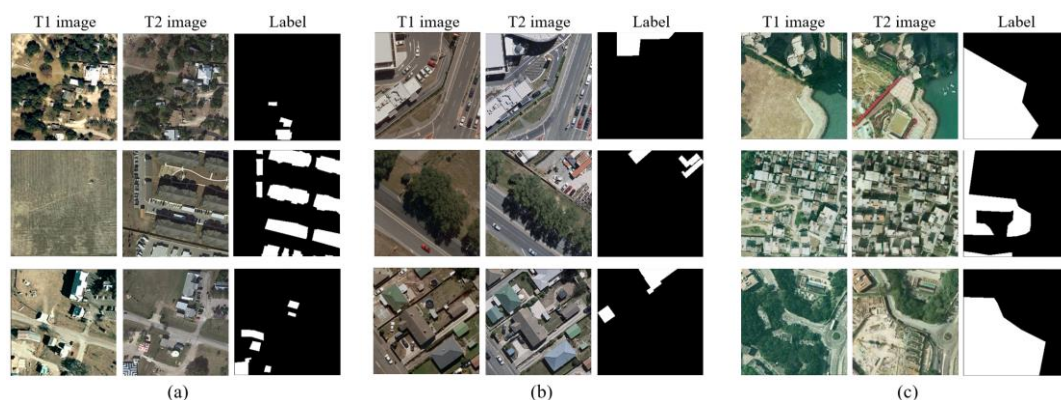


Figure 6 – (a) Samples from the LEVIR-CD dataset; (b) Samples from the WHU-CD dataset; (c) Samples from the SYSU-CD dataset.

3.3 Comparative experiments

3.3.1 Comparison algorithms

In this section, the proposed SPE-CD is compared with eight state-of-the-art methods, including five purely semi-supervised methods (S4GAN, SemiCDNet, SemiCD, CorrMatch and UniMatch), two methods that incorporate vision foundation models (SAM-CD and ChangeDINO) and one semi-supervised method built upon a foundation model (SemiCD-VL).

S4GAN [37]: This method introduces a generative adversarial network into semi-supervised CD, using adversarial learning between a generator and discriminator to exploit unlabelled data for improved feature discrimination.

SemiCDNet [15]: This method adopts consistency regularization with pseudo-label learning, enforcing prediction consistency under different perturbations to improve generalization from unlabelled data.

SemiCD [22]: This method builds a general semi-supervised framework based on pseudo-label supervision, leveraging unlabelled data to improve detection performance.

CorrMatch [38]: This method models pixel-level or region-level correspondences between bi-temporal features via correlation matching, improving change localization and feature consistency.

UniMatch [39]: This method integrates consistency learning with pseudo-labels under a unified semi-supervised framework, using multi-view augmentation to enhance unlabelled data utilization.

SAM-CD [40]: This method leverages the segmentation capability of SAM as an auxiliary module to improve boundary delineation and segmentation accuracy in CD.

ChangeDINO [31]: This method utilizes DINOv3 as a feature extractor and performs CD based on bi-temporal feature differencing under a semi-supervised setting.

SemiCD-VL [41]: This method incorporates a vision-language foundation model to introduce cross-modal semantic information, improving representation learning and robustness in complex scenes.

3.3.2 LEVIR-CD experiment results and analysis

The quantitative results on the LEVIR-CD dataset are reported in Table 2, including F1, IoU, precision and recall under different annotation ratios (5%, 10%, 20% and 40%). Overall, SPE-CD achieves the best F1 score and IoU across all settings, demonstrating the effectiveness of integrating semi-supervised learning with semantic priors from vision foundation models. Specifically, at 5% annotation, SPE-CD achieves an F1 of 91.73% and an IoU of 84.54%, outperforming the second best method (UniMatch) by 2.44% and 4.03%, respectively. At 10%, it further achieves 92.34% F1 and 85.84% IoU, exceeding UniMatch by 1.97% and 3.59%. At 20%, SPE-CD reaches 92.63% F1 and 86.36% IoU, surpassing SemiCD-VL by 1.25% and 2.27%. At 40%, it obtains 92.87% F1 and 86.69% IoU, outperforming ChangeDINO by 1.10% and 1.89%. Compared with the fully supervised baseline (Only-sup), SPE-CD consistently improves F1 by 3.41%, 2.21%, 1.57% and 1.36%, and IoU by 5.61%, 3.88%, 3.18% and 2.39% under increasing label ratios (5% to 40%). The performance gain gradually decreases as the annotation ratio increases, indicating that semantic priors from foundation models are particularly beneficial under low label conditions. Meanwhile, the marginal improvement of SPE-CD also gradually saturates with increasing labels, suggesting diminishing returns from additional supervision, while still maintaining a consistent advantage. Further analysis shows that under extremely low label settings (5%–10%), semi-supervised methods generally outperform methods relying solely on fully supervised training with foundation models, highlighting the importance of leveraging unlabelled data. As the annotation ratio increases, this gap narrows, and foundation model-based methods become increasingly competitive, reflecting their stronger ability to exploit semantic priors when supervision becomes moderately sufficient. Notably, SemiCD-VL achieves a consistently strong performance across all settings, further confirming the effectiveness of incorporating vision foundation model representations. In summary, SPE-CD consistently achieves superior performance across all annotation ratios, with particularly

significant gains in low label scenarios, demonstrating the effectiveness of combining semi-supervised learning with foundation model priors for CD.

Fig. 7 presents the prediction results of different methods on the LEVIR-CD dataset, where the label ratio used for training is 20%. To intuitively display the detection results, True Positives (TPs) are shown in white, True Negatives (TNs) in black, False Negatives (FNs, missed detections) in green and False Positives (FPs, false detections) in red. Specific regions in two of the examples are highlighted with local bounding boxes to better emphasize the details of the detection results. As shown in Fig. 7 (1)-(4), SPE-CD achieves the best overall detection performance compared with the other methods. Specifically, in Fig. 7 (1), SPE-CD provides the most complete detection of newly constructed buildings; under mild shadow and vegetation interference, it yields the fewest missed and false detections along the object edges. In Fig. 7 (2), SPE-CD achieves the most accurate detection of regular-scale buildings, with significantly reduced missed and false detections at the edges; in particular, it can still precisely recognize small-scale objects (as indicated by the blue box). In Fig. 7 (3), because the newly constructed buildings exhibit colours similar to the background, the compared methods, especially those without semantic prior information, tend to miss changes, thereby compromising the structural integrity of the objects. Although the compared methods that incorporate foundation models can detect changed regions more completely, their edge localization remains imprecise, resulting in numerous false detections (especially SAM-CD). By contrast, SPE-CD performs better, not only preserving the integrity of the objects but also effectively identifying small-scale change targets that the compared methods overlook. In Fig. 7 (4), for small-scale changed objects that are likewise difficult to distinguish from the background (highlighted by the blue boxes), the compared methods, particularly those without semantic prior information, often miss entire small-scale instances. Although methods that incorporate foundation models can detect the changed regions, the results remain incomplete. By comparison, SPE-CD still delivers the best performance, providing more complete and accurate detection of such targets. Overall, SPE-CD not only fully exploits the latent information contained in unlabelled data but also effectively integrates the semantic priors offered by the vision foundation model, thereby achieving the best visual performance on the building CD task.

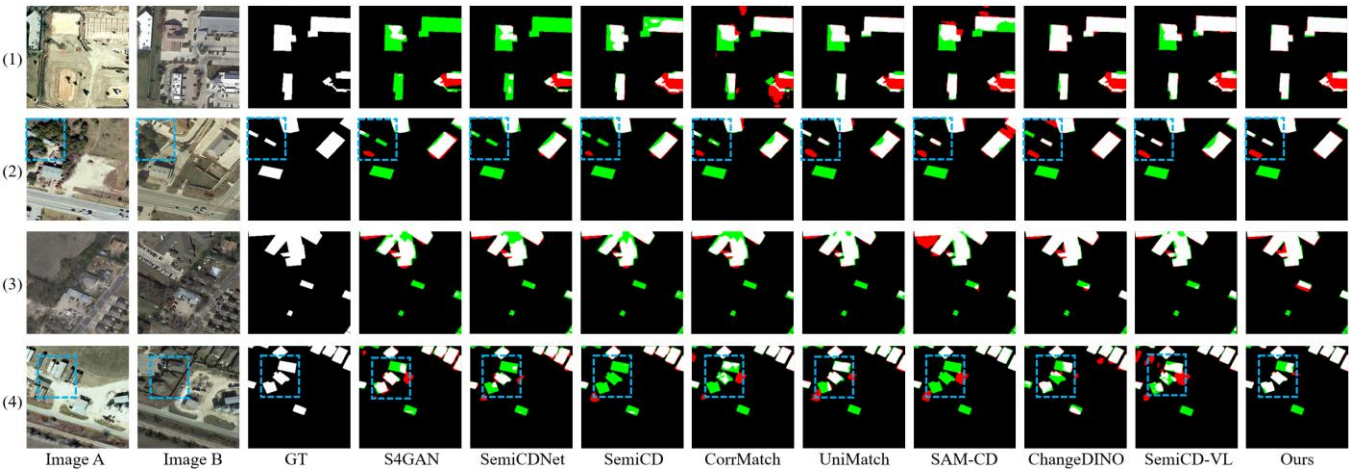


Figure 7 – Visualization results of different methods on the LEVIR-CD dataset under the 20% label ratio. (1)–(4) represent four sample groups from the test set.

Table 2 – Quantitative accuracy evaluation of different methods on the LEVIR-CD dataset.

Method	5%				10%			
	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)
Only-sup	88.32	78.93	89.33	87.32	90.13	81.96	91.24	89.05
S4GAN	78.26	64.29	86.70	71.33	81.57	68.93	87.96	76.01
SemiCDNet	81.52	68.81	88.60	75.49	84.26	72.84	89.47	79.63
SemiCD	85.18	74.19	91.37	79.79	87.18	77.24	91.62	83.26
CorrMatch	87.36	77.44	88.31	86.42	88.64	79.58	89.72	87.61
UniMatch	89.29	80.51	90.23	88.38	90.37	82.25	91.58	89.19
SAM-CD	83.91	72.18	86.83	81.27	85.86	75.36	88.34	83.62
ChangeDINO	86.74	76.42	88.05	85.51	88.96	80.13	90.26	87.82
SemiCD-VL	89.24	80.40	90.14	88.35	90.28	82.01	91.63	89.36
SPE-CD	91.73	84.54	92.70	90.79	92.34	85.84	91.58	93.13

Method	20%				40%			
	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)
Only-sup	91.06	83.18	91.03	91.09	91.51	84.30	92.30	90.75
S4GAN	84.08	72.56	88.79	80.16	86.78	76.64	89.48	84.24
SemiCDNet	85.79	75.14	90.27	81.82	86.47	76.16	90.15	83.07
SemiCD	87.91	78.46	91.12	85.21	88.26	78.99	90.83	85.84
CorrMatch	89.02	80.06	90.11	88.05	89.13	80.40	90.03	88.25
UniMatch	90.79	83.16	91.27	90.31	91.01	83.51	92.43	89.61
SAM-CD	88.87	80.14	91.08	86.81	90.76	83.17	92.30	89.36
ChangeDINO	91.19	83.86	92.51	90.04	91.77	84.80	92.71	90.85
SemiCD-VL	91.38	84.09	91.21	91.55	91.63	84.57	92.95	90.35
SPE-CD	92.63	86.36	92.88	91.42	92.87	86.69	93.00	92.75

3.3.3 WHU-CD experiment results and analysis

The quantitative results on the WHU-CD dataset are reported in Table 3, including F1, IoU, precision and recall under different annotation ratios (5%, 10%, 20% and 40%). Overall, SPE-CD achieves the best F1 and IoU across all settings, demonstrating the effectiveness of integrating semi-supervised learning with semantic priors from vision foundation models. Specifically, at 5% annotation, SPE-CD achieves an F1 of 92.56% and an IoU of 86.74%, outperforming the best competing methods by 2.70% and 3.89%, respectively, while also achieving 93.21% precision and 91.93% recall. At 10%, 20% and 40%, it improves F1 by 2.09%, 1.10% and 1.21%, and IoU by 3.68%, 1.99% and 0.24%, respectively, over the best competitors, while maintaining consistent superiority. Compared with the fully supervised baseline (Only-sup), SPE-CD improves F1 by 4.44%, 3.95%, 1.94% and 1.57%, and IoU by 3.89%, 3.68%, 2.73% and 1.20% under increasing label ratios, with gains gradually decreasing as supervision increases. Performance also remains stable across settings, indicating mild saturation at higher annotation ratios. Under low label conditions (5%–10%), semi-supervised methods outperform fully supervised foundation model-based approaches (e.g. SAM-CD). As the label ratio increases, methods such as ChangeDINO become competitive, showing improved utilization of semantic priors under sufficient supervision. SemiCD-VL also maintains a strong performance across all settings. In summary, SPE-CD consistently achieves superior performance, particularly under low label conditions, demonstrating the effectiveness of combining semi-supervised learning with foundation model priors for CD.

Fig. 8 presents the prediction results of different methods on the WHU-CD dataset, where the training label ratio is 20%. As shown in Fig. 8 (1)–(4), SPE-CD achieves the best overall performance compared with competing methods. In Fig. 8 (1), for irregularly shaped newly constructed buildings, semi-supervised methods tend to suffer from severe missed and false detections, while foundation model-based methods alleviate this issue to some extent but still struggle with accurate boundary localization. In contrast, SPE-CD produces more complete and precise change regions. In Fig. 8 (2), for small-scale irregular buildings, competing methods exhibit noticeable boundary errors, whereas SPE-CD achieves more accurate and fine-grained edge delineation. In Fig. 8 (3), in the presence of confusing objects such as vehicles, most competing methods fail to distinguish true changes from interference, resulting in significant missed detections. Although foundation model-based methods can capture more change regions, their results

remain incomplete. By comparison, SPE-CD shows stronger robustness and more complete change extraction under complex backgrounds. In Fig. 8 (4), for densely distributed irregular changes, competing methods often suffer from boundary blurring and object merging, whereas SPE-CD effectively preserves structural integrity and achieves more accurate boundary localization, leading to the best overall performance. Overall, SPE-CD effectively exploits unlabelled data while leveraging semantic priors from vision foundation models, resulting in superior visual performance for CD.

3.3.4 SYSU-CD experiment results and analysis

The quantitative results on the SYSU dataset are reported in Table 4, including F1, IoU, precision and recall under different annotation ratios (5%, 10%, 20% and 40%). Overall, SPE-CD achieves the best F1 and IoU across all settings, demonstrating the effectiveness of integrating semi-supervised learning with semantic priors from vision foundation models. Specifically, at 5% annotation, SPE-CD achieves an F1 of 81.36% and an IoU of 68.67%, outperforming the best competing method (UniMatch) by 1.48% and 2.24%, respectively. At 10%, 20% and 40%, it improves F1 by 0.98%, 0.19% and 0.54%, and IoU by 1.46%, 0.24% and 0.83%, respectively. Although the performance gap gradually narrows with increasing annotation ratios, SPE-CD consistently maintains the best performance. Compared with the fully supervised baseline (Only-sup), SPE-CD improves F1 by 5.15%, 3.80%, 2.49% and 1.28%, and IoU by 7.09%, 5.50%, 3.47% and 1.84% under 5%, 10%, 20% and 40% annotation ratios, respectively. The gains decrease as more labelled data is available, indicating that semantic priors are more beneficial under low label conditions. Meanwhile, performance improvements of SPE-CD itself gradually saturate with increasing annotations, suggesting diminishing returns from additional supervision while still maintaining overall superiority. Across methods, semi-supervised approaches generally outperform fully supervised methods incorporating foundation models under low label settings (5%–10%). As the annotation ratio increases, methods such as SAM-CD and ChangeDINO become more competitive, reflecting improved utilization of semantic priors under moderate supervision. SemiCD-VL consistently achieves a strong performance across all settings, further confirming the effectiveness of foundation model integration. In summary, SPE-CD consistently achieves superior performance across all annotation ratios, particularly under low label conditions, demonstrating the effectiveness of combining semi-supervised learning with foundation model priors for CD.

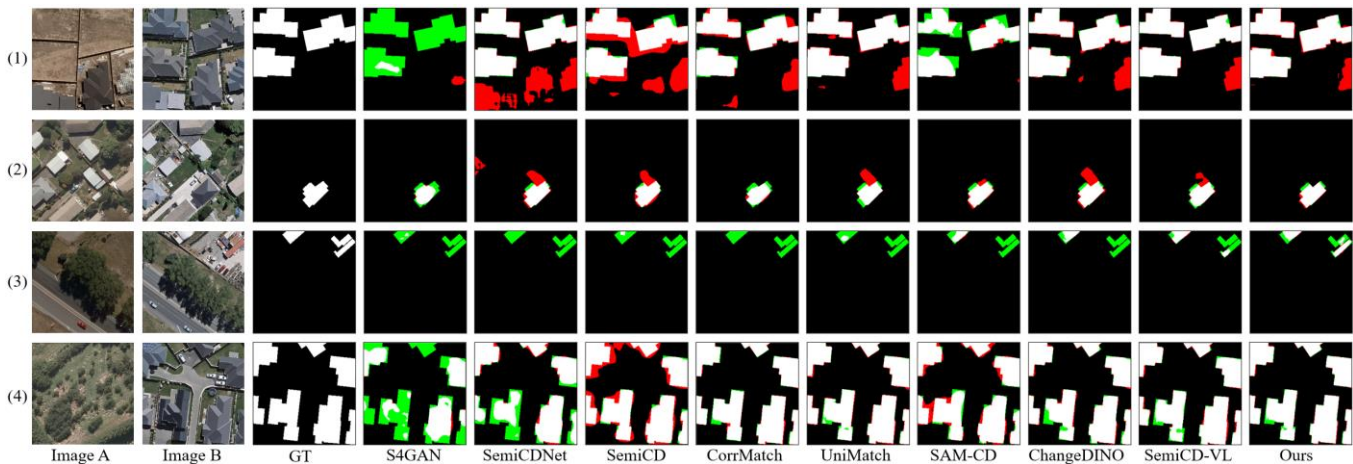


Figure 8 – Visualization results of different methods on the WHU-CD dataset under the 20% label ratio.

Table 3 – Quantitative accuracy evaluation of different methods on the WHU-CD dataset.

Method	5%				10%			
	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)
Only-sup	88.12	82.85	90.21	86.12	89.46	83.98	91.03	87.95
S4GAN	51.60	68.08	79.23	59.68	63.42	46.43	74.11	55.34
SemiCDNet	71.09	55.14	79.89	64.03	78.56	64.67	83.21	74.39
SemiCD	79.36	65.79	79.60	79.13	83.27	71.25	82.41	84.16
CorrMatch	85.18	75.23	88.37	82.21	87.91	78.45	90.12	85.80
UniMatch	89.86	82.13	93.19	86.76	91.32	83.92	93.21	89.53
SAM-CD	78.58	65.14	88.54	70.64	84.63	72.98	92.67	77.82
ChangeDINO	83.63	71.87	89.91	78.17	88.54	79.56	92.13	85.22
SemiCD-VL	89.70	81.32	94.00	85.77	91.08	83.61	93.55	88.74
Ours	92.56	86.74	93.21	91.93	93.41	87.66	94.02	92.81

Method	20%				40%			
	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)
Only-sup	91.27	84.62	92.54	90.03	92.18	85.96	93.12	91.26
S4GAN	80.15	67.04	85.41	75.50	86.25	75.83	91.20	81.81
SemiCDNet	84.67	73.45	88.02	81.57	88.32	79.08	90.15	86.55
SemiCD	85.94	75.36	86.73	85.17	87.65	78.01	86.67	88.65
CorrMatch	89.21	80.53	91.02	87.47	90.31	82.86	91.63	89.03
UniMatch	92.11	85.36	92.34	91.89	91.67	85.34	92.50	90.85
SAM-CD	90.93	82.10	93.93	88.11	90.86	80.31	91.71	90.03
ChangeDINO	91.83	84.82	93.22	90.48	92.54	86.92	93.64	91.47
SemiCD-VL	92.05	85.27	93.11	91.01	92.47	85.99	93.24	91.71
Ours	93.21	87.35	93.86	92.58	93.75	87.16	93.41	94.10

Table 4 – Quantitative accuracy evaluation of different methods on the SYSU-CD dataset.

Method	5%				10%			
	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)
Only-sup	76.21	61.58	77.84	74.66	78.94	65.10	80.12	77.79
S4GAN	68.72	52.36	77.15	61.96	71.43	55.51	78.94	65.21
SemiCDNet	70.18	54.06	78.23	63.66	72.96	57.36	80.02	66.90
SemiCD	73.41	58.01	81.52	66.78	70.61	54.54	89.14	58.61
CorrMatch	73.94	58.66	75.08	72.83	74.96	59.91	76.88	73.14
UniMatch	79.88	66.43	82.11	77.79	81.76	69.14	84.06	79.57
SAM-CD	73.85	58.57	76.92	71.03	76.18	61.47	79.63	73.02
ChangeDINO	78.52	64.46	80.96	76.24	80.91	67.92	83.11	78.85
SemiCD-VL	79.36	65.83	81.77	77.09	81.42	68.65	83.78	79.18
SPE-CD	81.36	68.67	83.02	79.76	82.74	70.60	84.15	81.37

Method	20%				40%			
	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)
Only-sup	80.63	67.61	81.74	79.55	82.34	70.01	83.45	81.28
S4GAN	72.94	57.38	83.12	64.86	75.16	60.23	86.37	66.52
SemiCDNet	72.36	56.64	80.74	65.57	74.32	59.13	82.29	67.75
SemiCD	72.52	56.88	91.03	60.26	75.86	61.18	83.40	69.55
CorrMatch	75.66	60.84	76.22	75.13	77.22	62.94	78.42	76.05
UniMatch	82.06	69.47	87.32	77.41	82.27	68.41	88.16	77.12
SAM-CD	76.51	61.96	82.05	71.68	78.12	64.10	82.14	74.48
ChangeDINO	81.17	68.24	84.83	77.68	82.21	69.70	84.92	79.69
SemiCD-VL	82.93	70.84	87.89	78.51	83.08	71.02	87.96	78.64
SPE-CD	83.12	71.08	87.12	78.29	83.62	71.85	88.21	79.08

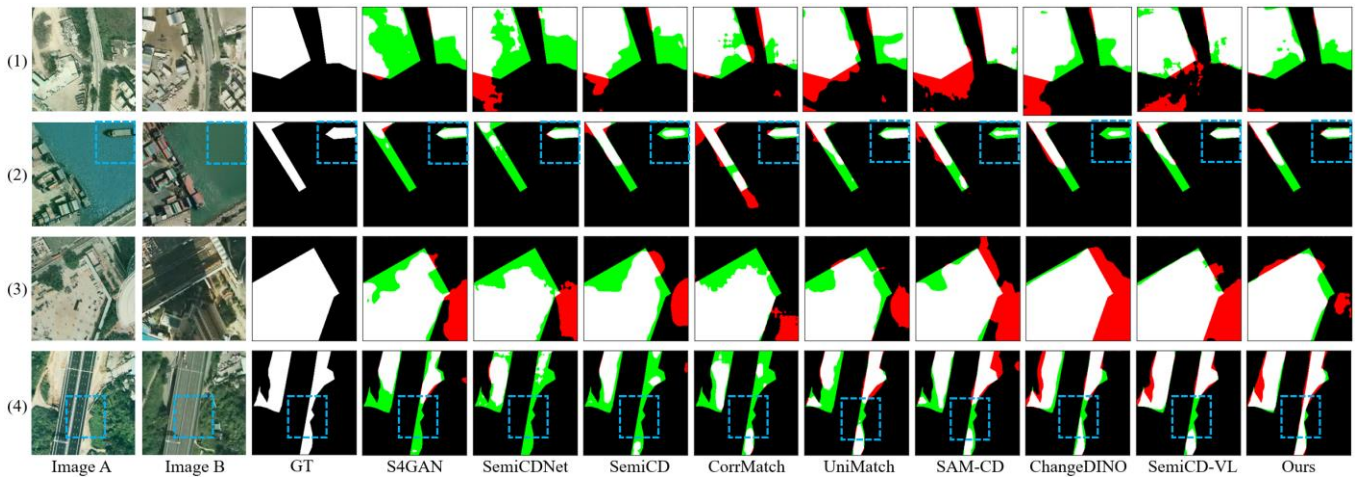
**Figure 9** – Visualization results of different methods on the SYSU-CD dataset under the 20% label ratio.

Fig. 9 presents the qualitative results of different methods on the SYSU-CD dataset, with a training label ratio of 20%. As shown in Fig. 9 (1)-(4), SPE-CD achieves the best overall performance compared with the competing methods. In Fig. 9 (1), for complex scenarios involving dense building changes and transitions between vegetation and bare land, SPE-CD more completely captures the changed regions, whereas the other methods suffer from obvious omissions, particularly failing to model fine-grained changes within dense building areas. In Fig. 9 (2), for newly added boats along water boundaries, purely semi-supervised methods tend to miss change objects due to insufficient semantic modelling, while foundation model-based methods can roughly locate changes but often produce false positives near boundary regions. In contrast, SPE-CD not only accurately localizes the changed area but also correctly identifies small-scale boats in the upper-right region (highlighted in the blue box), demonstrating stronger fine-detail perception. In Fig. 9 (3), under complex building-related change scenarios, semi-supervised methods struggle to fully cover the changed regions, whereas foundation model-based methods are prone to misclassifying background as change. SPE-CD effectively suppresses false positives while significantly reducing missed detections, yielding more reliable results. In Fig. 9(4), for changes between roadside bare land and vegetation, SPE-CD provides more accurate and spatially continuous boundary delineation. In particular, in local regions (blue box), most competing methods exhibit varying degrees of omission, whereas SPE-CD maintains better structural completeness. Overall, SPE-CD not only effectively exploits information from unlabelled data but also integrates semantic priors from vision foundation models, leading to more complete and accurate change region detection in complex scenes and consistently superior visual performance.

3.4 Ablation experiments

To comprehensively validate the effectiveness of each proposed module, we design systematic ablation experiments on the LEVIR-CD dataset. The training strategy is kept consistent across all models to ensure the fairness and comparability of the experimental results. The experimental design and the corresponding quantitative results are summarized in Table 5, where: (a) only the classical ResNet50 is used as the backbone for feature extraction, without the proposed SPE-Encoder; (b) the SPE-Encoder is adopted but the complete CCFP strategy is not used; instead, only cross-temporal feature perturbation is applied; (c) the SPE-Encoder is adopted, but the complete CCFP strategy is not used; instead, only the complementary feature perturbation strategy is applied; (d) the complete SPE-CD method.

(1) The SPE-Encoder substantially improves feature representation under low label conditions. As shown by the comparison between (a) and (d) in Table 5, removing the SPE-Encoder leads to a consistent performance degradation across all annotation ratios. At the 5% annotation ratio, the F1 score drops from 91.73% to 84.17% (−7.56%) and IoU from 84.54% to 72.36% (−12.18%). Similar trends are observed at 10%, 20% and 40%, where F1 decreases by 5.72%, 3.99% and 3.09%, and IoU by 10.59%, 7.03% and 5.95%, respectively. The performance gap gradually narrows as the annotation ratio increases, indicating that semantic priors are particularly critical under extreme label scarcity. In contrast to ResNet50-based feature extractors that rely heavily on limited supervision, the SPE-Encoder leverages pre-trained visual knowledge to provide more stable and discriminative semantic representations, effectively mitigating pseudo-label noise and improving semi-supervised CD performance.

Table 5 – Ablation results of each module on the LEVIR-CD dataset.

Method	SPE-Encoder	Complementary Perturbation	Cross-Temporal Perturbation	5%		10%	
				F1(%)	IoU(%)	F1(%)	IoU(%)
(a)	×	√	√	84.17	72.36	86.62	75.25
(b)	√	×	√	88.31	79.68	89.27	80.91
(c)	√	√	×	90.12	82.43	90.38	82.74
(d)	√	√	√	91.73	84.54	92.34	85.84

Method	SPE-Encoder	Complementary Perturbation	Cross-Temporal Perturbation	20%		40%	
				F1(%)	IoU(%)	F1(%)	IoU(%)
(a)	×	√	√	88.64	79.33	89.78	80.74
(b)	√	×	√	89.36	81.27	90.41	82.63
(c)	√	√	×	91.18	83.72	91.36	84.28
(d)	√	√	√	92.63	86.36	92.87	86.69

(2) The complementary feature perturbation strategy effectively enhances the diversity and robustness of feature representations. As shown by the comparison between (b) and (d) in Table 5, removing this module results in performance degradation across all annotation ratios. At 5% annotation, the F1 score decreases from 91.73% to 88.31% (−3.42%) and IoU from 84.54% to 79.68% (−4.86%). Similar declines are observed at 10%, 20% and 40%, with F1 dropping by 3.27%, 3.07% and 2.46% and IoU by 4.93%, 5.09% and 4.06%, respectively. These results show that constructing complementary feature subspaces enables the model to capture change information from multiple views, improving representation completeness. Compared with conventional random perturbation, the proposed structured channel masking reduces redundancy and enhances discriminative feature learning under complex change patterns.

(3) The cross-temporal feature perturbation strengthens the consistency modelling of change discrimination. As shown by the comparison between (c) and (d) in Table 5, its removal also leads to a performance drop across all annotation ratios. Specifically, F1 decreases by 1.61%, 1.96%, 1.45% and 1.51% at 5%, 10%, 20% and 40%, respectively, while IoU decreases by 2.11%, 3.10%, 2.64% and 2.41%. Although the performance drop is smaller than that caused by removing complementary perturbation, the improvement remains consistent. This is attributed to the temporal swap operation, which disrupts fixed temporal correspondence and encourages the model to learn invariant change representations under different temporal configurations, thereby improving robustness and generalization.

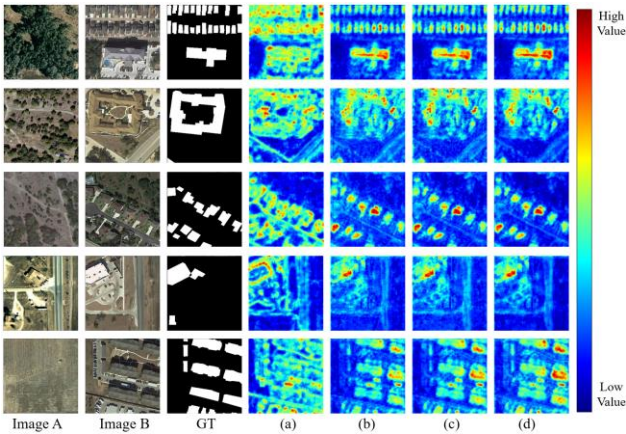


Figure 10 – Comparison of intermediate feature visualizations from the ablation experiments on the LEVIR-CD dataset under the 20% label ratio. (a) – (d) correspond to the different ablation settings, respectively.

To provide an intuitive comparison, Fig. 10 presents intermediate feature heatmaps of five randomly selected samples under the 20% training data ratio for different ablation settings, while Fig. 11 shows the corresponding CD results. From the feature visualizations, configuration (a), which does not include the SPE-Encoder, exhibits weak foreground responses, dispersed activations in changed regions, and limited contrast between foreground and background, leading to blurred boundaries. In addition, noticeable background activations indicate sensitivity to

texture noise, which degrades localization accuracy. Consistently, its CD results show incomplete structures, frequent missed detections, and coarse boundaries, particularly in small-scale or boundary-complex regions, highlighting insufficient feature discriminability. By contrast, configurations incorporating the SPE-Encoder (b, c, and d) show markedly stronger and more focused responses in changed regions, with improved background suppression. This demonstrates that semantic priors from pre-trained models enhance feature discriminability and improve separation between changed and unchanged areas. Further comparisons of perturbation strategies show that configuration (b), without complementary feature perturbation, produces relatively uniform responses but suffers from discontinuous or missing activations in some regions. Configuration (c), without cross-temporal perturbation, yields stronger overall responses but still shows false activations and incomplete boundaries in complex areas. In terms of CD results, (b) achieves reasonable coverage but exhibits fragmented predictions and local omissions, while (c) produces more complete regions but introduces additional false positives and boundary imprecision. These results suggest that single perturbation strategies cannot jointly ensure feature diversity and temporal consistency. In contrast, configuration (d) produces the most stable and well-structured responses, with strong activation in changed regions, clear boundaries, and effective suppression of background noise. Its CD results show more complete and continuous change regions, more accurate boundary localization, and significantly reduced missed and false detections, especially in small-scale and complex boundary cases. This confirms that the proposed CCFP strategy enhances feature diversity and improves robustness in modelling complex change patterns.

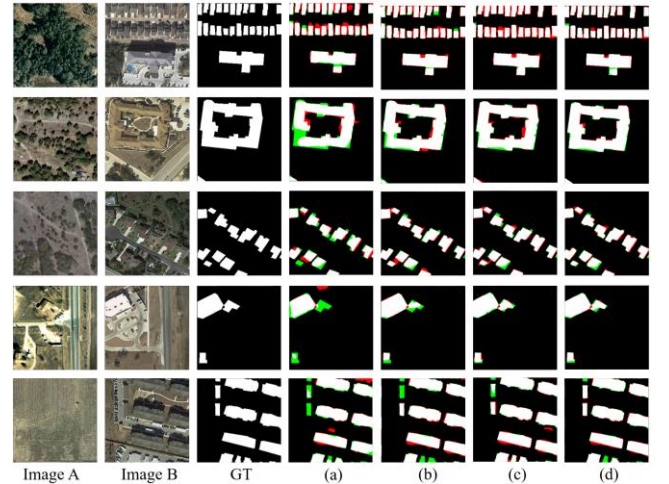


Figure 11 – Comparison of CD results from the ablation experiments on the LEVIR-CD dataset under the 20% label ratio. (a) – (d) correspond to the different ablation settings, respectively.

To further evaluate the spatial–frequency decoupled adaptation mechanism, ablation studies are conducted on the encoder using the LEVIR-CD dataset, considering three settings: (a) frozen backbone, (b) uniform fine-tuning, and (c) the proposed spatial–frequency decoupled adaptation. Results are reported in Table 6. Comparing (a) and (c),

freezing the backbone leads to a substantial performance drop across all annotation ratios. At 5% supervision, F1 decreases from 91.73% to 81.26% (−10.47%) and IoU from 84.54% to 69.18% (−15.36%). Similar trends are observed at higher ratios. This indicates that although pre-trained vision models provide strong semantic priors, direct transfer without adaptation fails due to the domain gap between natural images and remote sensing data, and may even cause negative transfer. Comparing (b) and (c), introducing fine-tuning significantly improves performance (e.g. F1 increases to 88.37% at 5%), confirming the necessity of adaptation. However, a performance gap remains, as uniform spatial-domain tuning ignores frequency-specific characteristics, limiting further gains. The gap is more pronounced under low label conditions, indicating that encoder design is particularly critical when supervision is scarce. Overall, the proposed spatial–frequency decoupled adaptation effectively mitigates cross-domain semantic shift and improves multi-granularity feature modelling, thereby enhancing robustness and accuracy in semi-supervised CD.

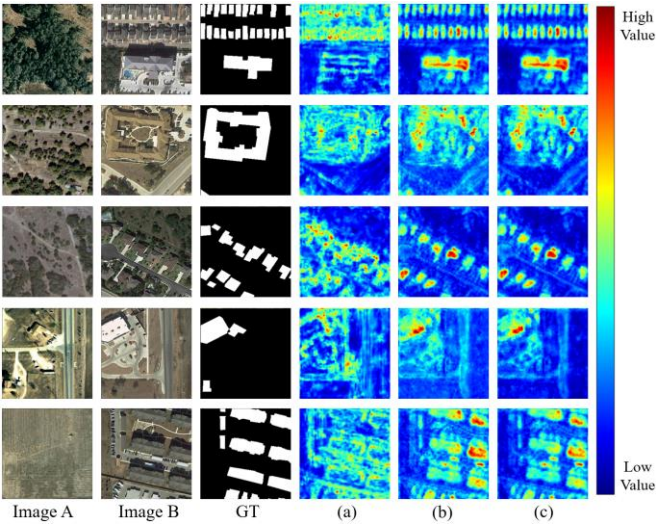


Figure 12 – Comparison of intermediate feature visualizations from the ablation experiments on the LEVIR-CD dataset under the 20% label ratio. (a) – (c) correspond to the different ablation settings, respectively.

Fig. 12 shows feature heatmaps under the same 20% training ratio for configurations (a), (b) and (c). Configuration (a) shows weak and poorly discriminative responses, reflecting poor transferability of frozen pre-trained features to remote sensing data. Configuration (b) improves foreground activation and overall feature quality but still exhibits blurred boundaries in regions with fine-grained changes due to uniform spatial fine-tuning. In contrast, configuration (c) produces the most structured and stable responses, with stronger and more continuous activations in changed

regions, better preserved structural integrity, and clearer boundary details. It also retains global consistency while enhancing local detail sensitivity, indicating effective integration of both low and high frequency information.

Table 6 – Ablation results for the encoder structure on the LEVIR-CD dataset.

Method	5%		10%	
	F1(%)	IoU(%)	F1(%)	IoU(%)
(a)Completely frozen	81.26	69.18	83.13	71.26
(b)Uniform fine-tuning	88.37	80.37	89.40	82.11
(c)Proposed (Ours)	91.73	84.54	92.34	85.84

Method	20%		40%	
	F1(%)	IoU(%)	F1(%)	IoU(%)
(a)Completely frozen	84.26	72.47	85.35	74.16
(b)Uniform fine-tuning	90.91	83.28	91.41	83.74
(c)Proposed (Ours)	92.63	86.36	92.87	86.69

3.5 Comparison of real-world change scenarios

To further verify the effectiveness and generalization of the proposed method in real-world settings, this section conducts supplementary experiments under realistic scenarios. Specifically, optical remote sensing images of Fushan District, Yantai City, Shandong Province, acquired by the Dongfang Huiyan Gaofen-01 satellite between 2024 and 2026, are used to construct a bi-temporal CD sample set that closely reflects practical application needs. The sample set focuses primarily on the dynamic changes of urban construction land, with the dominant change type being newly constructed buildings on bare land or vegetated areas, which is highly representative of real-world conditions. In terms of experimental setup, this section compares the proposed method with the mentioned eight competing methods. To ensure a fair comparison, all methods use the optimal model weights obtained by training on the LEVIR-CD dataset at the 40% annotation ratio. Given that the constructed real-world sample set is similar in change type to LEVIR-CD, where the dominant pattern is newly constructed buildings, these weights provide a strong basis for transfer. On this basis, all methods are directly applied to the real remote sensing data without any additional fine-tuning, allowing a unified evaluation of their robustness and generalization capability in real-world scenarios.

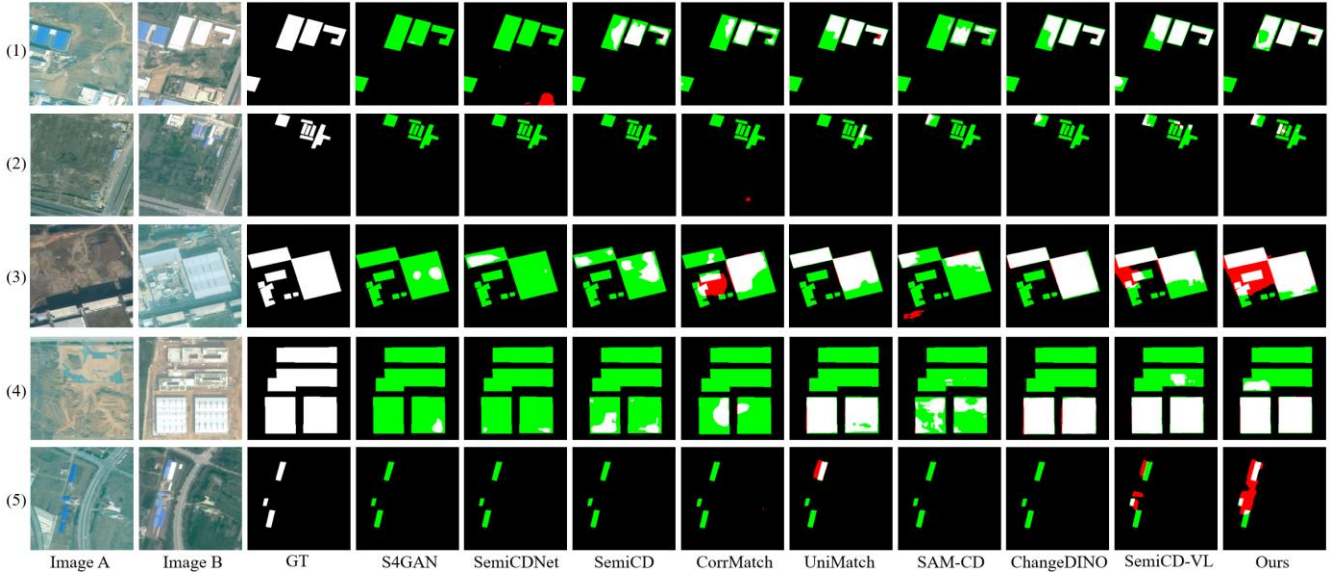


Figure 13 – Visualization results of different methods on the real-world sample set under the 40% label ratio. (1) – (5) represent the five sample groups in the sample set.

Table 7 – Quantitative results of different methods on the real-world samples.

Method	F1(%)	IoU(%)	Pre(%)	Rec(%)
S4GAN	15.60	8.46	99.86	8.46
SemiCDNet	18.36	10.11	98.01	10.12
SemiCD	30.38	17.91	99.17	17.94
CorrMatch	34.87	21.12	94.62	21.37
UniMatch	59.23	42.08	97.43	42.52
SAM-CD	54.79	37.73	97.21	38.14
ChangeDINO	63.86	46.91	99.12	47.08
SemiCD-VL	64.07	47.13	88.36	50.27
SPE-CD	65.27	48.44	80.26	54.96

The quantitative results on the five real remote sensing samples are reported in Table 7. Across all four evaluation metrics, SPE-CD achieves the best performance. Specifically, F1 and IoU reach 65.27% and 48.44%, exceeding the second best method, SemiCD-VL (64.07% and 47.13%), by 1.20% and 1.31%, respectively. In terms of Pre and Rec, SPE-CD also achieves a favourable balance, with Rec reaching 54.96%, which is markedly higher than that of the other methods, indicating a clear advantage in reducing missed detections. From the overall trend, purely semi-supervised methods (e.g. S4GAN and SemiCDNet) generally suffer from substantial missed detections and exhibit relatively low Rec values. In contrast, methods that incorporate foundation models (e.g. ChangeDINO and SemiCD-VL) can effectively enhance the perception of changed regions and substantially reduce the missed detection rate.

The corresponding qualitative visualization results are shown in Fig. 13. Overall, under few-shot training conditions, SPE-CD, which integrates the SPE-Encoder with an

effective semi-supervised training strategy, achieves superior detection performance on real-world samples and produces more complete and reliable CD results. Specifically, in Fig. 13 (1), all methods have some difficulty in identifying the newly constructed buildings in the upper right and fail to achieve complete detection. By comparison, SPE-CD identifies more detailed areas, with the smallest extent of missed detections and relatively more complete results. In Fig. 13 (2), the small-scale change targets further increase the detection difficulty, and the competing methods almost completely miss the changed objects, whereas SPE-CD is able to recover as many changed regions as possible. In Fig. 13 (3), the small-sized building cluster in the lower left is hard to recognize, and most methods miss it almost entirely. SPE-CD, despite producing some false detections, still manages to detect the newly constructed buildings as much as possible, which aligns more closely with the practical principle of "prioritizing the reduction of missed detections." In Fig. 13 (4), SPE-CD achieves the most complete detection of the large-scale buildings at the bottom, while also demonstrating a certain perception capability for the harder-to-recognize targets at the top, reflecting stronger generalization performance. In Fig. 13 (5), because the corresponding objects exhibit clear spatial misalignment, the boundary between changed regions and pseudo-changes becomes blurred, substantially increasing the detection difficulty. Although SPE-CD produces some false detections, it still covers more of the true changed regions, effectively reduces missed detections, and achieves the best overall detection performance. It should be noted that, although methods incorporating foundation models can learn richer semantic priors from large-scale data and thereby improve performance to a certain extent, they still exhibit limitations in real-world scenarios. This is mainly due to the distribution gap between real application data and training data, as well as the common presence of spatial misalignment, illumination variations, and complex background interference in real remote sensing imagery.

These factors make pseudo-changes harder to distinguish from real changes, thereby significantly increasing the overall difficulty of CD in practical applications.

4. CONCLUSION

In this article, we propose SPE-CD, a semantic prior-enhanced semi-supervised method for on-board change detection, designed to achieve robust CD under limited annotation conditions in high-resolution remote sensing imagery. To address the challenges posed by scarce labelled data and the underutilization of unlabelled data, SPE-CD effectively enhances feature representation and generalization by incorporating selectively fine-tuned semantic priors from a vision foundation model and a tailored feature perturbation mechanism. Specifically, an SPE-Encoder built upon a spatial-frequency decoupling mechanism adaptively corrects frequency band-wise domain shifts while preserving general semantic information, while the proposed CCFP strategy constructs complementary channel masks and cross-temporal feature swapping within a consistency learning framework to learn more stable and discriminative change features. Extensive experiments on three representative public datasets and real-world scenarios demonstrate that SPE-CD consistently outperforms several mainstream semi-supervised methods and state-of-the-art foundation model-based approaches in both quantitative and qualitative results, highlighting its strong robustness to complex change patterns. In the future, we plan to further reduce the model's dependency on manually labelled data through more efficient foundation model adaptation strategies and lightweight semi-supervised learning frameworks.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62371352 and 62425102. The authors gratefully acknowledge Star.ai Spatio-temporal Intelligence Technology Co., Ltd. for providing the datasets and technical support.

REFERENCES

- [1] Ashbindu Singh. "Review article digital change detection techniques using remotely-sensed data". In: *International Journal of Remote Sensing* 10.6 (1989), pp. 989-1003 doi: 10.1080/01431168908903939.
- [2] Adam Van Etten, Daniel Hogan, Jesus Martinez-Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. "The multi-temporal urban development SpaceNet dataset". In: *arXiv preprint arXiv:2102.04420* (2021) doi: 10.48550/arXiv.2102.04420 url: <http://arxiv.org/abs/2102.04420>.
- [3] Mengxi Liu, Zhuoqun Chai, Haojun Deng, and Rong Liu. "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 4297-4306 doi: 10.1109/JSTARS.2022.3177235.
- [4] Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters". In: *Remote Sensing of Environment* 265 (2021), pp. 112636 doi: 10.1016/j.rse.2021.112636.
- [5] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. "xBD: a dataset for assessing building damage from satellite imagery". In: *arXiv preprint arXiv:1911.09296* (2019) doi: 10.48550/arXiv.1911.09296 url: <http://arxiv.org/abs/1911.09296>.
- [6] Chahira Serief, Youcef Ghelamallah, and Youcef Bentoutou. "Deep-learning-based system for change detection onboard earth observation small satellites". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), pp. 8115-8124 doi: 10.1109/JSTARS.2023.3284919.
- [7] Gabriele Inzerillo, Diego Valsesia, Aniello Fiengo, and Enrico Magli. "Compress-align-detect: onboard change detection from unregistered images". In: *arXiv preprint arXiv:2507.15578* (2025) doi: 10.48550/arXiv.2507.15578 url: <http://arxiv.org/abs/2507.15578>.
- [8] Vít Růžička, Anna Vaughan, Daniele De Martini, James Fulton, Valentina Salvatelli, Chris Bridges, Gonzalo Mateo-Garcia, and Valentina Zantedeschi. "RaVÆn: Unsupervised change detection of extreme events using ML on-board satellites". In: *Scientific Reports* 12.1 (2022), pp. 16939 doi: 10.1038/s41598-022-19437-5.
- [9] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. "Fully convolutional siamese networks for change detection". In: *2018 25th IEEE International Conference on Image Processing (ICIP)* IEEE. 2018, pp. 4063-4067 doi: 10.1109/ICIP.2018.8451652 url: <https://ieeexplore.ieee.org/document/8451652/>.
- [10] Yang Zhan, Kun Fu, Menglong Yan, Xian Sun, Hongqi Wang, and Xiaosong Qiu. "Change detection based on deep siamese convolutional network for optical aerial images". In: *IEEE Geoscience and Remote Sensing Letters* 14.10 (2017), pp. 1845-1849 doi: 10.1109/LGRS.2017.2738149.

- [11] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 166 (2020), pp. 183-200 doi: 10.1016/j.isprsjprs.2020.06.003.
- [12] Hao Chen, Zipeng Qi, and Zhenwei Shi. "Remote sensing image change detection with transformers". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1-14 doi: 10.1109/TGRS.2021.3095166.
- [13] Xiaofeng Zhang, Shuli Cheng, Liejun Wang, and Haojin Li. "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection". In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1-15 doi: 10.1109/TGRS.2023.3245674.
- [14] Wele Gedara Chaminda Bandara, and Vishal M. Patel. "A transformer-based siamese network for change detection". In: *IGARSS 2022 – 2022 IEEE International Geoscience and Remote Sensing Symposium* 2022, pp. 207-210 doi: 10.1109/IGARSS46834.2022.9883686 url: <https://ieeexplore.ieee.org/document/9883686/>.
- [15] Daifeng Peng, Lorenzo Bruzzone, Yongjun Zhang, Haiyan Guan, Haiyong Ding, and Xu Huang. "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.7 (2021), pp. 5891-5906 doi: 10.1109/TGRS.2020.3011913.
- [16] Jia-Xin Wang, Teng Li, Si-Bao Chen, Jin Tang, Bin Luo, and Richard C. Wilson. "Reliable contrastive learning for semi-supervised change detection in remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1-13 doi: 10.1109/TGRS.2022.3228016.
- [17] Xueting Zhang, Xin Huang, and Jiayi Li. "Semisupervised change detection with feature-prediction alignment". In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1-16 doi: 10.1109/TGRS.2023.3247605.
- [18] Qidi Shu, Jun Pan, Zhuoer Zhang, and Mi Wang. "MTCNet: multitask consistency network with single temporal supervision for semi-supervised building change detection". In: *International Journal of Applied Earth Observation and Geoinformation* 115 (2022), pp. 103110 doi: 10.1016/j.jag.2022.103110.
- [19] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. "Domain generalization: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 1-20 doi: 10.1109/TPAMI.2022.3195549.
- [20] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. "FixMatch: simplifying semi-supervised learning with consistency and confidence". In: *arXiv preprint arXiv:2001.07685* (2020) doi: 10.48550/arXiv.2001.07685 url: <http://arxiv.org/abs/2001.07685>.
- [21] Ran Lingyan, Wen Dongcheng, Zhuo Tao, Zhang Shizhou, Zhang Xiuwei, and Zhang Yanning. "AdaSemiCD: An adaptive semi-supervised change detection method based on pseudo-label evaluation". In: *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025), pp. 1-14 doi: 10.1109/TGRS.2025.3551504.
- [22] Wele Gedara Chaminda Bandara, and Vishal M. Patel. "Revisiting consistency regularization for semi-supervised change detection in remote sensing images". In: *arXiv preprint arXiv:2204.08454* (2022) doi: 10.48550/arXiv.2204.08454 url: <http://arxiv.org/abs/2204.08454>.
- [23] Wuxia Zhang, Xinlong Shu, Siyuan Wu, and Songtao Ding. "Semi-supervised change detection with data augmentation and adaptive thresholding for high-resolution remote sensing images". In: *Remote Sensing* 17.2 (2025), pp. 178 doi: 10.3390/rs17020178.
- [24] Chaoning Zhang, Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Chenghao Li, Yu Qiao, Taegoo Kang, Xinru Shan, Chenshuang Zhang, Caiyan Qin, Francois Rameau, Lik-Hang Lee, Sung-Ho Bae, and Choong Seon Hong. "A survey on segment anything model (SAM): vision foundation model meets prompt engineering". In: *arXiv preprint arXiv:2306.06211* (2024) doi: 10.48550/arXiv.2306.06211 url: <http://arxiv.org/abs/2306.06211>.
- [25] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. "Vision models are more robust and fair when pretrained on uncurated images without supervision". In: *arXiv preprint arXiv:2202.08360* (2022) doi: 10.48550/arXiv.2202.08360 url: <http://arxiv.org/abs/2202.08360>.
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning transferable visual models from natural language supervision". In: *arXiv preprint arXiv:2103.00020* (2021) doi: 10.48550/arXiv.2103.00020 url: <http://arxiv.org/abs/2103.00020>.
- [27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. "Segment anything". In: *arXiv preprint arXiv:2304.02643* (2023) doi: 10.48550/arXiv.2304.02643 url: <http://arxiv.org/abs/2304.02643>.

- [28] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection". In: *arXiv preprint arXiv:2203.03605* (2022) doi: 10.48550/arXiv.2203.03605 url: <http://arxiv.org/abs/2203.03605>.
- [29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. "DINOv2: learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2024) doi: 10.48550/arXiv.2304.07193 url: <http://arxiv.org/abs/2304.07193>.
- [30] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jegou, Patrick Labatut, and Piotr Bojanowski. "DINOv3". In: *arXiv preprint arXiv:2508.10104* (2025) doi: 10.48550/arXiv.2508.10104 url: <http://arxiv.org/abs/2508.10104>.
- [31] Yixin Chen, Ruiqian Zhang, Xiaogang Ning, Hanchao Zhang, You He, Yuxing Xie, and Jiaming Wang. "Change dino: A unified transformer-based framework for object-level change detection and segmentation in remote sensing imagery". In: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium 2024*, pp. 8585-8589 doi: 10.1109/IGARSS53475.2024.10642342 url: <https://ieeexplore.ieee.org/document/10642342/>.
- [32] Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. "ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 208 (2024), pp. 53-69 doi: 10.1016/j.isprsjprs.2024.01.004.
- [33] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. "Randaugment: practical automated data augmentation with a reduced search space". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* IEEE. 2020, pp. 3008-3017 doi: 10.1109/CVPRW50498.2020.00359 url: <https://ieeexplore.ieee.org/document/9150790/>.
- [34] Hao Chen, and Zhenwei Shi. "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection". In: *Remote Sensing* 12.10 (2020), pp. 1662 doi: 10.3390/rs12101662.
- [35] Shunping Ji, Shiqing Wei, and Meng Lu. "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.1 (2019), pp. 574-586 doi: 10.1109/TGRS.2018.2858817.
- [36] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1-16 doi: 10.1109/TGRS.2021.3085870.
- [37] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. "Semi-supervised semantic segmentation with high- and low-level consistency". In: *arXiv preprint arXiv:1908.05724* (2019) doi: 10.48550/arXiv.1908.05724 url: <http://arxiv.org/abs/1908.05724>.
- [38] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. "CorrMatch: Label propagation via correlation matching for semi-supervised semantic segmentation". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE. 2024, pp. 3097-3107 doi: 10.1109/CVPR52733.2024.00299 url: <https://ieeexplore.ieee.org/document/10657798/>.
- [39] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation". In: *arXiv preprint arXiv:2208.09910* (2023) doi: 10.48550/arXiv.2208.09910 url: <http://arxiv.org/abs/2208.09910>.
- [40] Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. "Adapting segment anything model for change detection in VHR remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1-11 doi: 10.1109/TGRS.2024.3368168.
- [41] Kaiyu Li, Xiangyong Cao, Yupeng Deng, Jiayi Song, Junmin Liu, Deyu Meng, and Zhi Wang. "SemiCD-VL: visual-language model guidance makes better semi-supervised change detector". In: *arXiv preprint arXiv:2405.04788* (2024) doi: 10.48550/arXiv.2405.04788 url: <http://arxiv.org/abs/2405.04788>.

AUTHORS



KEXIN ZHU received a BS degree from Nanjing Normal University, Nanjing, China, in 2023. She is currently working towards a master's degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. Her research interests include remote sensing change detection.



ZIZHEN ZHANG received a BS degree from Hohai University, Nanjing, China, in 2025. He is currently working towards a master's degree in resources and environment with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include remote sensing change detection.



JUN PAN received a B.Eng. degree in geographic information science and M.Sc. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002, 2005, and 2008, respectively. Since 2018, he has been a professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. His research interests include remote sensing image intelligent processing, data fusion, image reconstruction, image restoration and image quality improvement.