S^3CD: A Self-Supervised Semantic Change Detection Method by Mining Transition Patterns and Consistency in Remote Sensing Images

Yang Qu, Jiayi Li, *Senior Member, IEEE*, Xiaofeng Pan, and Xin Huang, *Fellow, IEEE*

***Abstract*—Semantic change detection (SCD) endeavors to identify land-cover changes from multitemporal remote sensing images, providing essential information for various applications. Nevertheless, conventional supervised SCD methods necessitate extensive pixel-level annotations, limiting their applicability. The capability of self-supervised methods to learn feature representations with large amounts of unlabeled data and minimal annotation, and to achieve superior performance, has made them one of the hot topics in remote sensing. However, most self-supervised methods in remote sensing are primarily designed to learn general semantic representations of images, which limits their effectiveness for tasks like SCD that require the analysis of complex semantic transformations. To address this, we propose a multistage, multitask, and multilevel self-supervised network, named S^3CD, that learns semantic changes from bi-temporal remote sensing images across scene, pixel, and prototype levels in two stages. In particular, in Stage 2, the network enhances the robustness of SCD by learning semantic consistency within the semantic stable categories across different temporal and capturing the temporal patterns of semantic change categories. We evaluate S^3CD on two widely used remote sensing change detection (CD) datasets, where it outperformed state-of-the-art self-supervised and supervised SCD methods. Notably, in the binary CD (BCD) task (i.e., detecting the locations of changes), S^3CD also outperforms most supervised learning methods. Therefore, this approach facilitates the application of self-supervised learning in the field of remote sensing CD.**

***Index Terms*—Multilevel, multistage, remote sensing semantic change detection (SCD), self-supervised.**

I. Introduction

IN RECENT years, increasing urbanization and industrialization have heightened interest in understanding the interactions between human activities and land surface changes. To address this, change detection (CD) techniques have been developed to identify, extract, and analyze change targets using multitemporal remote sensing images from the same area [1]. With advancements in remote sensing platforms, CD techniques are now widely applied in fields, such as forest CD [2], building expansion monitoring [3], and farmland change analysis [4].

Popular CD methods focus on identifying changed pixel locations in multitemporal remote sensing images, i.e., binary CD (BCD) [5], [6]. However, many practical applications require not only accurate localization but also detailed descriptions of the change type. To solve this problem, researchers have proposed the concept of semantic CD (SCD) [7], [8], [9]. In contrast to BCD, SCD is more informative, using bi-temporal images to generate both a binary change map and a “from-to” land-cover change map indicating the direction of change. This technology not only identifies changed areas (achieving the BCD task), but also links changes to specific land-cover categories, providing a more comprehensive understanding of land-cover processes. SCD thus offers detailed transition (i.e., “from-to”) information for practical applications like urban planning, environmental monitoring, and disaster surveillance.

Deep learning methods have made significant advancements in the field of image processing and analysis [10], [11], [12]. In particular, convolutional neural networks (CNNs), which are capable of analyzing the spatio-temporal dependencies inherent in multitemporal remotely sensed images, have been widely applied to SCD tasks. However, most state-of-the-art SCD methods rely on fully supervised frameworks, requiring extensive pixel-level annotations of bi-temporal images. This annotation requirement poses significant practical constraints, as manual labeling is both resource-intensive and time-consuming. While transfer learning and pretraining strategies have been proposed to leverage knowledge from auxiliary datasets [13], [14], [15], their effectiveness may be compromised when there are significant domain gaps between the source and target datasets.

Recently, self-supervised learning methods [16], [17] have shown potential for SCD by extracting generic feature representations from a large amount of unlabeled image data relevant to the target task. These features are then transferred to downstream tasks, enabling superior performance even with limited labeled samples. This approach has proven effective in various remote sensing applications, including semantic segmentation [18], CD [19], and scene classification [20]. Research suggests that customizing self-supervised networks

Received 23 July 2025; revised 21 October 2025; accepted 23 October 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFF1306102 and Grant 2022YFB3903701 and in part by the National Natural Science Foundation of China under Grant 42471391 and Grant 42271328. This article was recommended by Associate Editor J. Park. *(Corresponding author: Jiayi Li.)*

Yang Qu, Jiayi Li, and Xin Huang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: quyang@whu.edu.cn; zjjerica@whu.edu.cn; xhuang@whu.edu.cn).

Xiaofeng Pan is with Shenzhen Ecological and Environmental Monitoring Center Station of Guangdong Province, Shenzhen 518049, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2025.3625757.

Digital Object Identifier 10.1109/TCYB.2025.3625757

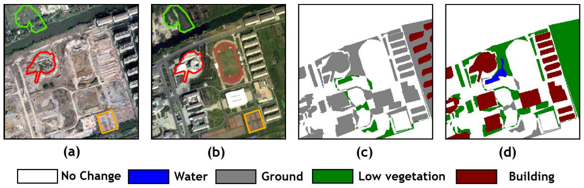


Fig. 1. SECOND [8] dataset (a) time-1 ($T1$) image, (b) time-2 ($T2$) image, (c) land-cover map for the semantic changed region at $T1$, and (d) land-cover map for the semantic changed region at $T2$.

for specific downstream tasks can yield significant advantages [21], [22]. However, the application of remote sensing SCD remains limited, mainly due to two challenges:

1) *Limitations in Capturing Semantic Transition Patterns:* Existing remote sensing self-supervised CD algorithms can learn temporal-invariant semantic representations and identify whether land-cover/land-use (LCLU) changes occur. However, these methods overlook the varying temporal transition patterns between semantic categories in bi-temporal remote sensing images. For example, in Fig. 1(c) and (d), the difference between the "from ground to low vegetation" changes and the "from ground to building" changes was not explored in previous CD methods. This limitation restricts the network's ability to capture both the temporal dynamics within semantic categories and the distinct transition patterns between categories, ultimately reducing their effectiveness in SCD tasks that involve complex transitions across multiple LCLU types.

2) *Incorrect Positive Samples in Self-Supervised CD:* Existing self-supervised remote sensing CD algorithms typically leverage images from the same location taken at different times as positive samples to learn semantic representations of LCLU through temporal changes. However, bi-temporal images can exhibit various types of changes: natural temporal changes [green polygons of Fig. 1(a) and (b)], semantic changes caused by human activities [red polygons in Fig. 1(a) and (b)], and state changes that do not alter the semantic [orange rectangles of Fig. 1(a) and (b)]. Existing self-supervised methods often assume that there is no semantic change in the scene, thereby learning semantic invariant information. However, these approaches often inevitably take semantic changes caused by human activities as unchanged semantic information. That is, it leads to the issue of "incorrect" self-supervised samples and hinders the effective mining of semantic features, especially in pixel-based self-supervised methods for SCD.

To tackle the challenges mentioned above, we propose S^{3}CD, a self-supervised framework designed to learn semantic change information from bi-temporal remote sensing images during pretraining. It is a multistage, multitask, and multilevel self-supervised SCD framework. In the first stage, the model is encouraged to learn coarse semantic information by minimizing the similarity between different augmented views of a single image and maximizing the dissimilarity between different images. In the second stage, a multitask framework is designed to guide the learning of the semantic temporal-invariance (i.e., the semantic consistency of the same categories at different times) and the semantic change pattern from the bi-temporal images.

In summary, the main contributions of this study, which are in the second stage of the proposed network, are as follows.

1) A multitask learning framework is designed to improve the S^{3}CD network's ability to leverage semantic information from bi-temporal images. This framework decomposes bi-temporal SCD into three complementary subtasks: modeling fine-grained semantic temporal-invariance, capturing semantic change patterns, and enhancing the distinction between change and stable features. Unlike existing self-supervised methods that treat CD as a single task, this approach captures fine-grained semantic change information at multiple levels (i.e., prototypes and pixels), reducing task complexity and enhancing overall performance.

2) The semantic alignment (SA) module is proposed to model fine-grained semantic temporal-invariance. The module is realized by minimizing the distance between the corresponding prototypes in bi-temporal feature space and maximizing interprototype distances in the single-temporal feature space. This dual optimization achieves more efficient feature recognition and semantic representation learning.

3) The contrastive change pattern (CCP) module is proposed to capture the semantic transition pattern. The module enhances the sensitivity of the model to semantic changes in land-cover transition by aligning prototypes with the same semantic change patterns and amplifying the differences between prototypes of semantic change patterns. CCP integrates Dual-Mamba to model temporal dependencies in semantic change, thus achieving robust capture of complex time dynamics.

The subsequent sections of this article are structured as follows. Section II offers a review of pertinent literature concerning SCD and self-supervised methods. Section III expounds on the S^{3}CD self-supervised method. Section IV details the experimental setup and evaluation metrics. Section V presents the experimental results. Then, further discussion are carried out in Section VI. Finally, Section VII concludes this work and provides conclusive remarks.

II. Related Work

A. Deep Learning-Based SCD

BCD is designed to distinguish between changed and unchanged pixels by outputting a binary change map [23]. However, multiple change types are included in practical CD applications, including urban expansion [24] and deforestation [25]. To achieve a more comprehensive understanding of these changes, the notion of SCD has been introduced in studies [26]. It can characterize land-cover type shifts between bi-temporal images, providing important transition information that effectively addresses the questions of "where" and "how" the changes have occurred. The intuitive solution to SCD is to consider each "from-to" change as a category, and then concatenate and input the bi-temporal remote sensing images into a semantic segmentation model to generate a "from-to" change type for each pixel [27], [28]. However, this approach requires modeling each possible semantic change type, leading to a significant increase in the label space and requiring a large amount of labeled data for effective network training. Consequently, postclassification comparison (PCC) methods have emerged as a prevalent alternative for SCD tasks [29].

Traditional PCC methods use semantic segmentation models to generate a separate semantic map for each image, which are then compared to detect changes [30]. However, these maps fail to capture the intraclass correlation of the same ground objects across different temporal images and are prone to cumulative errors. In addition, they do not fully leverage temporal correlation, which is crucial for CD, thus limiting their ability to discriminate change features. To address these limitations, recent research has explored SCD methods based on a multitask framework, using a Siamese semantic segmentation model to extract bi-temporal land-cover and change features while concurrently optimizing both segmentation and CD tasks [31], [32]. However, the aforementioned methods still rely on fully supervised deep learning models, which require a large number of carefully annotated samples, significantly restricting the practical applicability of supervised learning-based SCD methods.

Furthermore, it is important to note that SCD tasks face another significant challenge: the annotation process is time-consuming and labor-intensive. To facilitate the annotation of large-scale SCD datasets, a common practice is to uniformly label the majority of unchanged pixels in the dataset as "unchanged," only annotating the changed LCLU types. Therefore, in order to improve computational efficiency and adapt to this marking scheme, most SCD models predominantly focus on extracting changed region semantic change information while neglecting the different semantic representation of the same LCLU in bi-temporal images. This limits the model's capacity to comprehensively characterize and identify semantic changes. This limits the model's capacity to comprehensively characterize and identify semantic changes.

B. Self-Supervised Learning

The primary limitation of previous supervised SCD methods is their limited generalizability to practical applications, largely due to the extensive need for labeled samples. Furthermore, a substantial amount of unlabeled remote sensing data remains underutilized. In this context, the ongoing development of self-supervised learning techniques, which can automatically extract intrinsic features from unlabeled data [33], [34], [35], [36], has paved the way for novel self-supervised BCD methods, demonstrating considerable potential [37].

Self-supervised BCD methods typically consist of two phases: pretraining and fine-tuning. In the pretraining phase, a self-supervised framework is designed to guide the network in learning the invariance of unlabeled remote sensing image transformations by enforcing similarity between different augmented views of the same image (positive sample pairs) and dissimilarity between different images (negative sample pairs). Subsequently, in the fine-tuning phase, the pretrained model is trained using a small number of labeled samples to achieve performance comparable to supervised methods [38].

Recent studies have demonstrated that leveraging bi-temporal remote sensing image pairs directly as multiple views in self-supervised BCD tasks can effectively facilitate the learning of temporal-invariance [39]. While this approach provides more realistic temporal representations, it assumes that minimal or no semantic changes occur between image pairs. However, this assumption is not always valid. For example, commonly used semantic change datasets like the SECOND dataset [8], around 10% pixels exhibit changes, which is not negligible. In the pixel-level feature learning process, such a proportion of semantic change pixels can considerably degrade model training. To mitigate this issue, some studies have attempted to employ pseudo-change labels to extract bi-temporal image patches with high similarity as positive sample pairs [40], [41]. Although this strategy reduces the interference of change pixels, it still ignores the various LCLU transitions information between the bi-temporal images [42], [43].

In conclusion, the direct application of existing self-supervised BCD methods to SCD tasks may fail to fully capture temporal changes between bi-temporal images and may be affected by semantic changes, potentially degrading model performance. To address the aforementioned issues, this article proposes a self-supervised SCD network, aiming to learn the semantic change information through two stages in the pretraining phase. In the first stage, the semantic information of a single image is learned through different views of the image. In the second stage, the semantic feature maps of bi-temporal images are compared to explore the semantic information between the semantic change and the semantic stable regions, thereby gaining a deeper understanding of the semantic details of land features at various levels in bi-temporal images.

III. Methodology

A. Overview

In this work, we propose a novel self-supervised framework specifically designed for SCD. As illustrated in Fig. 2, our approach employs a two-stage self-supervised architecture with a shared backbone network. First, in the first Stage 1, focus on learning single-temporal semantic features. Then, in Stage 2, multilevel semantic representations are modeled across bi-temporal image pairs. Specifically, Stage 1 (the upper in Fig. 2): learning coarse semantic features from a single-temporal image. At this stage, different augmented views of a single-temporal image are used as positive pairs, while other images serve as negative samples. Stage 2 (the bottom in Fig. 2): capturing fine-grained semantic change information from bi-temporal images. At this stage, inspired by multitask learning, a popular supervised SCD paradigm, the proposed self-supervised version is decomposed into subtasks of modeling fine-grained semantic temporal-invariance, capturing semantic change patterns, and enhancing differences between change and stable features.

B. Stage 1 (Learning Coarse Semantic Features From a Single-Temporal Image)

It aims to maximize the similarity between positive pairs (i.e., different augmented views of the same remote sensing

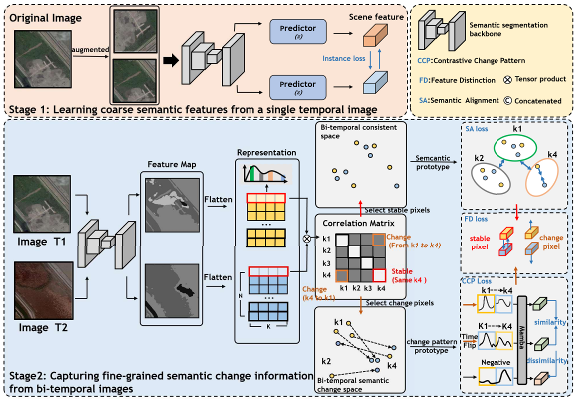


Fig. 2. Proposed S^{3}CD framework. Stage 1, augmented views of a single-temporal image are taken as inputs, with instance loss guiding the model to learn scene-level semantic representation. Stage 2, the semantic segmentation backbone processes bi-temporal images to generate feature maps, each of which is then flattened along spatial dimensions. The rows of each feature map represent pixel probabilities for K clusters, and N is the number of pixels in the image pair. A bi-temporal correlation matrix is computed via tensor products of spatially aligned features, capturing semantic transitions between the two time points. By applying the argmax operation along both rows and columns of the correlation matrix, the semantic stable and change features are identified, resulting in two complementary feature subspaces: the bi-temporal semantic-stable and semantic-change spaces. Then, SA and CCP are used to mine the semantic consistency and the semantic change pattern, respectively. Finally, FD losses are computed to enhance the distinction between change and stable features, thereby facilitating the learning of SA and CCP.

image) and minimize the similarity between negative pairs in order to facilitate the model can focus on learning single-temporal scene-level semantic information. Following the standard contrastive learning framework [44], multiple data augmentation techniques are applied to generate diverse views, while other samples in the same batch serve as negatives. In this study, HRNet_w40 [45] is used as a backbone; the feature maps output by the backbone are of the same size as the input image. These features are then further transformed through a nonlinear projection head, implemented as a multilayer perceptron (MLP) and an adaptive mean pooling operation [44]. Finally, the learning process is guided by the $\mathcal{L}_{ins}$ (i.e., InforNCE loss function [44]).

C. Stage 2 (Capturing Fine-Grained Semantic Change Information From Bi-Temporal Images)

In remote sensing imagery, bi-temporal image pairs often stem from semantic-irrelevant factors, such as seasonal variations or satellite viewing geometry. Therefore, using bi-temporal images of pixels as positive sample pairs in Stage 2 helps the model learn semantic temporal-invariance. However, it should be noted that there are still some semantic changes caused by human activities. Consequently, directly using corresponding pixels from bi-temporal images as positive samples could introduce erroneous training information.

To address this issue, bi-temporal features are adaptively partitioned into two parts: semantic stable features (i.e., state change but semantic stable) and semantic change features. To be specific, as shown in Fig. 2, after the first stage of training, the backbone extracts features from the input bi-temporal images as $\mathcal{F}_{T1} \in \mathbb{R}^{N\times K}$ and $\mathcal{F}_{T2}$, where N represents the total number of pixels, and K is the dimensional of the extracted features. Subsequently, through the tensor product operation of the pixels at both time points, a correlation matrix characterizing the semantic transition relationship can be obtained. Within this matrix framework, rows and columns, respectively, correspond to the semantic states of pixels at $T1$ and $T2$. Applying the argmax operation along both rows and columns of the correlation matrix reveals the pixel-wise semantic relationships between $T1$ and $T2$. Then, the semantic stable features $\hat{\mathcal{F}}_{T1} \in \mathbb{R}^{M\times K}$ and $\hat{\mathcal{F}}_{T2}$, and semantic change features $\widetilde{\mathcal{F}}_{T1} \in \mathbb{R}^{L\times K}$ and $\widetilde{\mathcal{F}}_{T2}$ are distinguished, where M represents the total number of pixels in semantic stable feature, L represents the total number of pixels in semantic change feature, and $M + L = N$.

The complex task of learning semantic change information is then decomposed into three interrelated subtasks: 1) modeling fine-grained semantic temporal-invariance (prototype-based); 2) capturing semantic change patterns (prototype-based); and 3) enhancing distinctions between change and stable features (pixel-based). The remaining part of Section III provides detailed descriptions of these subtasks.

1) *Subtask (a): Modeling Fine-Grained Semantic Temporal-Invariance:*

Full supervised SCD approaches typically treat unchanged regions as background. While these regions may not exhibit significant changes, they still contain valuable semantic information that can enhance model performance. To leverage this untapped potential, a SA module is designed. This module simultaneously maximizes prototype separability within individual temporal feature spaces and minimizes the distance between corresponding bi-temporal prototypes in semantically stable regions, enabling the model to learn semantic temporal invariance for more effective feature discrimination. To improve efficiency, we adopt an online clustering strategy that avoids the high complexity $O(NK^2)$ of global methods like k-means, achieving linear time complexity of $O(NK)$.

Let, $\hat{\mathcal{F}}_{T1}$ and $\hat{\mathcal{F}}_{T2}$ note bi-temporal semantic stable feature, where each rows of this feature matrix $\hat{\mathcal{F}}_{T1}$ represent the probability distribution of pixels across K cluster, while the columns correspond to the probability of M pixel belonging to a specific cluster (i.e., clustering representation) [46]. Based on this, the prototypes can be calculated

$$\mu_{T1}^{k} = \frac{\sum_{m=1}^{M} p\left(k|m_{T1}\right) \hat{\mathcal{F}}_{T1}^{m}}{\left\| \sum_{m=1}^{M} p\left(k|m_{T1}\right) \hat{\mathcal{F}}_{T1}^{m} \right\|_2} \tag{1}$$

where $k \in [1, K]$, μ_{T1}^{k} denotes kth prototype in $T1$, $p\left(k \mid |m_{T1}\right)$ is the probability that mth pixel is assigned to kth prototype at $T1$. $\hat{\mathcal{F}}_{T1}^{m}$ represents the feature vector of mth pixel in $\hat{\mathcal{F}}_{T1}$.

In the context of bi-temporal semantic stable features, the number of prototypes remains consistent across different temporal. Therefore, the matching prototype across time from positive pairs is $\{\hat{\mu}_{T1}^{k}, \hat{\mu}_{T2}^{k}\}$, and all others as negatives. The similarity between prototypes can thus be quantified as

$$\begin{aligned} s\left(\hat{\mu}_{T1}^{k}, \hat{\mu}_{T2}^{k}\right) &= \frac{\left(\hat{\mu}_{T1}^{k}\right)^{\top}\left(\hat{\mu}_{T2}^{k}\right)}{\tau_k} \\ &= \sum_{m=1}^{M} \frac{\left(p\left(k|m_{T1}\right) \hat{\mathcal{F}}_{T1}^{m}\right)^{\top}\left(p\left(k|m_{T2}\right) \hat{\mathcal{F}}_{T2}^{m}\right)}{\left\|p\left(k|m_{T1}\right) \hat{\mathcal{F}}_{T1}^{m}\right\|_2 \left\|p\left(k|m_{T2}\right) \hat{\mathcal{F}}_{T2}^{m}\right\|_2 \tau_k} \end{aligned}$$

$$= \sum_{m=1}^{M} \underbrace{\frac{p\left(k|m_{T1}\right) p\left(k|m_{T2}\right)}{\left\| p\left(k|m_{T1}\right) \hat{F}_{T1}^{m} \right\|_2 \left\| p\left(k|m_{T2}\right) \hat{F}_{T2}^{m} \right\|_2}}_{\text{semantic alignment}} \underbrace{\frac{\left(\hat{\mathcal{F}}_{T1}^{m}\right)^{\top} \left(\hat{\mathcal{F}}_{T2}^{m}\right)}{\tau_k}}_{\text{feature alignment}} \quad (2)$$

where τ_k is a temperature parameter. Without loss of generality, the following loss function is employed to differentiate all prototypes other than μ_{T1}^{k} and μ_{T2}^{k}. The expression is as follows:

$$\ell_{T1}^{k} = -\log \frac{\exp\left(s\left(\hat{\mu}_{T1}^{k}, \hat{\mu}_{T2}^{k}\right)\right)}{\exp\left(s\left(\hat{\mu}_{T1}^{k}, \hat{\mu}_{T2}^{k}\right)\right) + \sum_{\substack{j=1 \\ j\neq k}}^{K} \exp\left(s\left(\hat{\mu}_{T1}^{k}, \hat{\mu}_{T1}^{j}\right)\right)}. \quad (3)$$

Since samples within a mini-batch may not cover all prototypes, the loss and logits of the empty prototype are zeroed out in each iteration. By traversing all prototypes in $\hat{\mathcal{F}}_{T1}$ and $\hat{\mathcal{F}}_{T2}$, the SA loss is calculated as

$$\mathcal{L}_{SA} = \frac{1}{2K} \sum_{k=1}^{K} \left(\ell_{T1}^{k} + \ell_{T2}^{k}\right) - \mathrm{H}\left(\mathcal{F}_{T1}, \mathcal{F}_{T2}\right) \quad (4)$$

where $H(\mathcal{F}_{T1}, \mathcal{F}_{T2})$ is the entropy of prototype assignment probabilities to ensure a balanced distribution of prototype assignments [47].

2) *Subtask (b): Capturing Semantic Change Patterns:* The fundamental objective of SCD is to accurately identify and land-cover transition patterns in change regions, which is essential for understanding land-cover dynamics. Therefore, the CCP module is proposed to capture semantic change patterns in semantic change regions by aligning prototypes with the same semantic change patterns and amplifying the differences between prototypes of distinct semantic change patterns. However, a fundamental challenge in the approach is constructing a positive sample pair that exhibits a consistent semantic change pattern. To address this, we propose a novel method for generating positive sample pairs by leveraging the forward and reverse within semantic change patterns. Specifically, flipping the temporal order of reverse change so that they align with the pattern of forward changes, thus ensuring the consistency of semantic changes. This method cannot only automatically generate complementary sample pairs from existing semantic change patterns, but also enhance the ability of the model to explain the dynamics of semantic evolution.

Formally, to capture semantic change patterns, we concatenate bi-temporal semantic change features $\widetilde{\mathcal{F}}_{T1}$ and $\widetilde{\mathcal{F}}_{T2}$ along the feature dimension to form $\widetilde{\mathcal{F}} \in \mathbb{R}^{L\times 2K}$, where L represents the total number of pixels in the semantic change feature. The semantic change space can theoretically accommodate $K^2 - K$ different change patterns, including forward and reverse change (assuming that a change pattern from prototype $i \rightarrow j$ is defined as forward if $i < j$, and reverse if $i > j$, where $i, j \in [1, K]$). Then, the rows and columns of change correlation matrix $\widetilde{Y} \in \mathbb{R}^{L\times K\times K}$ (from the tensor product of corresponding pixels at $\widetilde{\mathcal{F}}_{T1}$ and $\widetilde{\mathcal{F}}_{T2}$) can represent the probability of pixel-wise semantic change. Therefore, the prototype of change patterns is $\{v_{1,2}, v_{1,3}, \ldots, v_{K,K-1}\}$

$$v_{i,j} = \frac{\sum_{l=1}^{L} \tilde{Y}^{l,i,j} \tilde{\mathcal{F}}^{l}}{\left\| \sum_{l=1}^{L} \tilde{Y}^{l,i,j} \tilde{\mathcal{F}}^{l} \right\|_2} \quad (5)$$

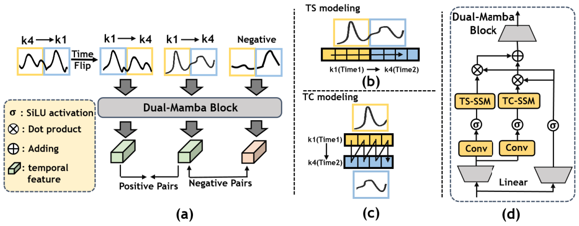


Fig. 3. (a) Schematic of semantic change pattern extraction, where $k1, k4 \in [1, K]$. The reverse change pattern after time flip and the forward change pattern are input into the Dual-Mamba network to generate positive feature pairs. Another change pattern in the Dual-Mamba network output as negative pairs. (b) TS modeling, with bi-temporal features arranged chronologically to capture the chronologically changing pattern. (c) TC modeling, interleaving time features to capture the dependencies between bi-temporal. (d) Dual-Mamba block architecture. The input features are processed through two parallel linear layers (two branches). The first branch consists of two subpaths with a 1-D convolution layer with SiLU activation and an SSM. The TS-SSM and TC-SSM represent two variants of SSM that integrate TC and TS modeling strategies, respectively.

where $v_{i,j}$ represents the prototype of semantic change from i to j, $\widetilde{Y}^{l,i,j}$ is the value of the change correlation matrix of lth pixel at row i and column j. $\widetilde{\mathcal{F}}^{l}$ denotes feature of lth pixel in $\widetilde{\mathcal{F}}$.

Then, by flipping the feature connection order of the reverse change pattern, $S = ((K^2 - K)/2)$, positive prototype pairs of the change pattern with the same change direction can be formed $\{(\overrightarrow{v}_1, \overleftarrow{v}_1), (\overrightarrow{v}_2, \overleftarrow{v}_2), \ldots, (\overrightarrow{v}_S, \overleftarrow{v}_S)\}$, where $(\overrightarrow{v}_S, \overleftarrow{v}_S)$ represents Sth positive prototype pair, $\overrightarrow{v}$ denotes the a prototype with forward change, and $\overleftarrow{v}$ represents the a prototype with forward change after flipping.

Meanwhile, simply concatenating bi-temporal semantic change features may not fully capture the temporal dynamics in semantic change. To address this limitation, we introduce Dual-Mamba networks, which combine the Mamba network [48] designed for time series data analysis, with a complementary modeling scheme that integrates temporal sequential (TS) and temporal cross (TC) modeling schemes (see Fig. 3).

Finally, the model is guided to simulate and capture the semantic change pattern by maximizing the similarity of the same semantic change pattern, to enhance the discernibility between different change patterns

$$\mathcal{L}_{CCP} = \sum_{s=1}^{S} \Bigg(\underbrace{s\left(\vec{f}_s, \overleftarrow{f}_s\right)}_{\text{prototype alignment}} + \underbrace{\log \sum_{\substack{a=1 \\ a\neq s}}^{S} \exp\left(s\left(\vec{f}_s, \vec{f}_a\right)\right)\Bigg)}_{\text{prototype uniformity}} \quad (6)$$

where $\{\overrightarrow{f}_s, \overleftarrow{f}_s\}$ represents the output of sth prototype pair through mamba. As not all change patterns appear in mini-batches, we zero out both the loss terms and logit values for each iteration involving empty clusters and no matching bidirectional change.

3) *Multitask Combination (Enhancing Differences Between Change and Stable Features):* To further enhance the model's ability to discriminate between semantic change and stability, we introduce a feature distinction (FD) module. While subtask (a) learns temporal-invariance and subtask (b) models change patterns, both benefit from clearer separation between stable and change

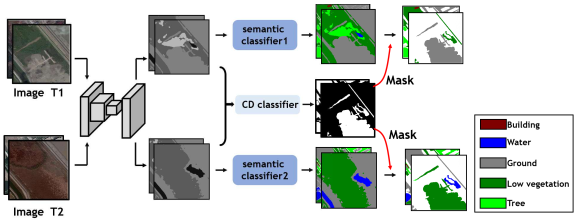


Fig. 4. Fine-tuning framework of S^{3}CD. The fine-tuning framework for the SCD task comprises a pretrained shared-weight backbone, two semantic classifiers, and a CD classifier.

semantics. FD strengthens this separation by maximizing the distance between change features while minimizing the distance between stable features.

Specifically, the feature distance between bi-temporal images is calculated to generate a change intensity map. Subsequently, a change threshold is adaptively determined by identifying the "corner" on the distribution curve of the change intensity map and computing its maximum deviation from the straight line connecting the endpoints of the curve [19]. This threshold is then applied to segment the change intensity map, producing a pseudo-change map that separates change features from stable features. Finally, the semantic change and stable features obtained from subtasks (a) and (b) are intersected with the outputs of the above method to derive high-confidence changes and stable features. Based on this, a positive pair $\{\mathcal{F}_{T1}^{+}, \mathcal{F}_{T2}^{+}\}$ and a negative pair $\{\mathcal{F}_{T1}^{-}, \mathcal{F}_{T2}^{-}\}$ are constructed, where $\{\mathcal{F}_{T1}^{+}, \mathcal{F}_{T2}^{+}\}$ represents high-confidence semantic stable feature pairs, $\{\mathcal{F}_{T1}^{-}, \mathcal{F}_{T2}^{-}\}$ represents high-confidence semantic change feature matrix pairs. The pixel-level FD loss is computed as

$$\mathcal{L}_{FD} = \frac{1}{B}\sum_{i=1}^{B} 1 - \left(\mathrm{sim}\left(\mathcal{F}_{T1}^{i+}, \mathcal{F}_{T2}^{i+}\right) - \mathrm{sim}\left(\mathcal{F}_{T1}^{i-}, \mathcal{F}_{T2}^{i-}\right)\right) \tag{7}$$

where sim refers to cosine similarity, and B is the mini-batch size.

D. Overall Loss

The self-supervised pretraining process for S^{3}CD consists of two stages, $\mathcal{S}_1$ and $\mathcal{S}_2$ epochs. For $\mathcal{S}_1$ (Stage 1), $\mathcal{L}_{\text{ins}}$ is used to update the model weights. For $\mathcal{S}_2$ (Stage 2), we design a multitask loss combination to guide the model to expand from scene-level feature representation to pixel-level and prototype-level, as shown in the following equation:

$$\mathcal{L}_{Stage2} = \lambda\mathcal{L}_{FD} + (1-\lambda)\left(\mathcal{L}_{SA} + \mathcal{L}_{\text{CCP}}\right) \tag{8}$$

where λ controls the balance between SA, CCP, and FD.

To determine specific "from-to" land-cover transitions, the pretrained network requires fine-tuning with semantic annotations, including LULC and change location labels. As illustrated in Fig. 4. In fine-tuning process of S^{3}CD, we design two loss functions: semantic class loss $\mathcal{L}_{seg}$, and binary change loss $\mathcal{L}_{change}$. Specifically, the semantic loss $\mathcal{L}_{seg}$ is the multiclass cross-entropy between predicted semantic segmentation and the ground-truth semantic change map. The $\mathcal{L}_{seg}$ of each pixel is calculated as follows:

$$\mathcal{L}_{seg} = -\frac{1}{C}\sum_{c=1}^{C} y_c \log\left(p_c\right) \tag{9}$$

where C is the number of reference coverage categories in the dataset, and y_c and p_c denote the predicted probability of the reference label and ith class, respectively. It is worth noting that unchanged categories (indexed as "0") are excluded from the loss computation to encourage the semantic branch to focus on extracting semantic features. The change loss $\mathcal{L}_{change}$ is the binary cross-entropy loss between the predicted binary change map $\hat{y}_c$ and the reference label y_c. The $\mathcal{L}_{change}$ for each pixel is computed as

$$L_{change} = -y_c \log\left(p_c\right) - \left(1-y_c\right)\log\left(1-p_c\right). \tag{10}$$

The training of the two semantic branches is directly supervised by the semantic segmentation map, while the training of the CD classifier is directly supervised by $\mathcal{L}_{change}$. The overall loss $\mathcal{L}_{scd}$ for the SCD task is denoted as

$$\mathcal{L}_{scd} = \left(\mathcal{L}_{seg_1} + \mathcal{L}_{seg_2}\right)/2 + \mathcal{L}_{change} \tag{11}$$

where $\mathcal{L}_{seg_1}$ and $\mathcal{L}_{seg_2}$ are the semantic losses of the two temporal image branches.

IV. Data Description and Experimental Setup

A. Dataset

This study evaluates the performance of the model primarily on three open-source datasets, described as follows.

1) *SECOND Dataset [8]:* This dataset comprises 4662 pairs of bi-temporal high-resolution images, with 2968 pairs being publicly available. The publicly accessible data was randomly divided into training and testing sets in a 9:1 ratio [9]. The images in these pairs have consistent dimensions of 512 × 512 pixels, with spatial resolutions ranging from 0.5 to 3 m. They encompass six land-cover categories.
2) *HRSCD Dataset [7]:* Comprising 291 pairs of aerial images, each image in this dataset is of size 10 000 × 10 000 pixels with a spatial resolution of 0.5 m. The dataset covers five land-cover categories. Following the strategy outlined in [9], semantic change maps were generated based on land-cover maps and binary change maps to align with the SECOND dataset. Subsequently, the original images were cropped into nonoverlapping 512 × 512 image pairs using a sliding window. The training and testing sets were randomly split in a 9:1 ratio.
3) *Hi-UCD Mini Dataset [49]:* Comprises multitemporal aerial imagery acquired over three consecutive years (2017–2019) in Tallinn, Estonia. Each image has dimensions of 1024 × 1024 pixels with a fine ground sampling distance of 0.1 m. The dataset categorizes land-cover into ten distinct classes. As in [50], all images were cropped into 512 × 512 patches, and unchanged image pairs were excluded from the analysis. This preprocessing yields 571 training, 100 validation, and 705 testing image pairs.

B. Comparative Algorithms for SCD

To comprehensively validate the performance of our proposed self-supervised SCD model, we selected several state-of-the-art supervised SCD networks: HRSCD.str3 [7], HRSCD.str4 [7], HRNet-based SCD (HBSCD), SSCD-l [32], Bi-SRNet [32], SCanNet [51], MCTNet [52], HGINet [53], and CDLNet [54]. All competing methods are experimented with based on the descriptions and optimal hyperparameter settings in the original paper.

C. Evaluation Metrics for SCD

In semantic change tasks, the highly imbalanced dataset labels lead to the fact that commonly used evaluation metrics for BCD tasks (e.g., overall accuracy) can easily be dominated by negative samples, which do not provide reasonable accuracy metrics. Therefore, we follow the existing setup in the literature [8] to use mean intersection over union (mIoU) and separation Kappa coefficient (SeK) to evaluate the model's performance. And, the overall score (Score) of the model can be calculated based on mIoU and Sek

$$\text{Score} = 0.3 * \text{mIoU} + 0.7 * \text{Sek}. \tag{12}$$

D. Experimental Setup

All experiments in this study were conducted using the PyTorch library on an NVIDIA RTX2080Ti GPU. The proposed S^{3}CD method employs HRNet-w40 as the backbone framework. For the pretraining phase of all self-supervised methods, the stochastic gradient descent (SGD) algorithm was utilized for 50 epochs on the target dataset, with a batch size of 20, a learning rate of 0.03, weight decay of 1e^{-4}, and momentum of 0.9. Models requiring momentum updates had their momentum update parameter set to 0.998. The best models obtained during the pretraining phase for self-supervised methods were retained for downstream tasks. Specifically, S^{3}CD consists of two stages, with 45 and 5 epochs, and τ_k and τ_l set to 1 and 0.5, respectively (referenced from [46]).

To ensure fairness, all self-supervised methods in either the fine-tuning phases used the same experimental parameters, including a batch size of 3, optimization method (SGD), epochs (30), and initial learning rate (0.0015). All methods utilized pretrained parameters from ImageNet.

V. Results

In this section, we first comparatively evaluate the performance of S^{3}CD on the SECOND, HRSCD, and Hi-UCD mini datasets. Next, we provide a visual analysis of the bi-temporal semantic feature representations learned by the pretrained S^{3}CD model. Through ablation experiments, we examine the individual contributions of each component of S^{3}CD. Finally, investigating the impact of Mamba structure on semantic change modeling.

A. Experimental Results and Analysis

To intuitively display the SCD results of various methods on each dataset, we selected different scenes from each dataset for visualization, as shown in Fig. 5. It can be observed that, in target-dense scenarios, S^{3}CD not only detects small-sized change areas more effectively than other models but also captures finer details and edges. It is noteworthy that methods like BiSRNet and HBSCD optimize two semantic segmentation heads using only changing pixels. This configuration often undermines the network's ability to leverage labeled data, resulting in limited CD performance. In contrast, S^{3}CD fully utilizes the semantic information of unchanged pixels during the pretraining phase, and this is applied to the target task, significantly enhancing recognition capabilities. In addition, Fig. 5 clearly illustrates that most models struggle to identify the "from water to artificial surface" change type in the HRSCD dataset, while our method accurately recognizes this change category.

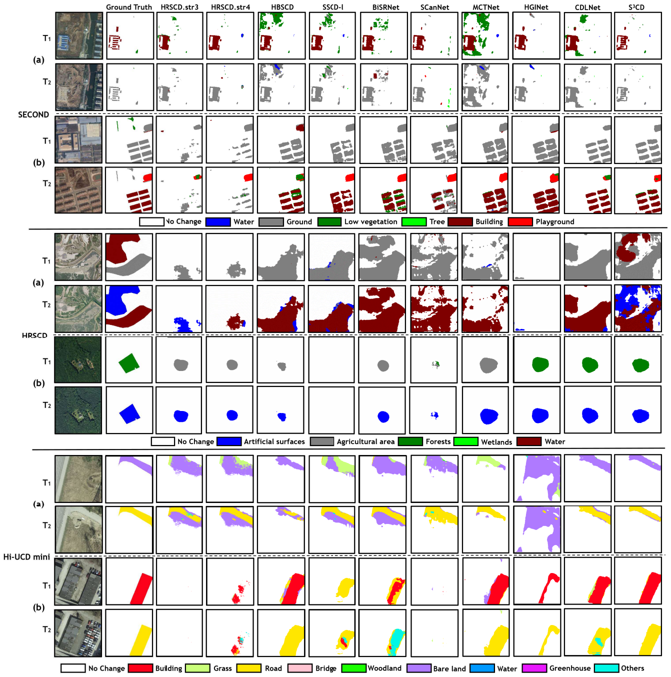


Fig. 5. Visual results of S^{3}CD with the SOTA method on the different datasets. (a) and (b) SCD results obtained from three pairs of different bi-temporal images.

TABLE I
Experimental Results Comparing the SOTA Supervised Method and S^{3}CD Method for SECOND Dataset

Model	Computation Costs		Accuracy		
	Params(M)	FLOPs(G)	Score(%)	mIoU(%)	Sek(%)
HRSCD-st.3	12.77	42.94	26.17	64.48	9.74
HRSCD-st.4	13.71	43.97	33.42	70.44	17.55
HBSCD	49.10	175.84	35.88	71.28	20.70
SSCD-l	23.31	189.76	35.85	71.64	20.51
BiSRNet	23.39	189.91	36.23	71.90	20.94
SCanNet	27.90	264.95	36.55	72.01	21.35
MCTNet	26.36	76.71	36.19	71.57	21.04
HGINet	27.70	50.63	36.50	71.54	21.48
CDLNet	12.88	30.10	37.00	72.21	21.90
S^{3}CD	46.27	129.17	39.07	73.58	24.28

Tables I–III present the results of various methods on the different datasets. Consistent with the visual results, the S^{3}CD model outperforms all other methods in terms of Score, mIoU,

TABLE II

EXPERIMENTAL RESULTS COMPARING THE SOTA SUPERVISED METHOD AND S^{3}CD METHOD FOR HRSCD DATASET

Model	Computation Costs		Accuracy		
	Params(M)	FLOPs(G)	Score(%)	mIoU(%)	Sek(%)
HRSCD-st.3	12.77	42.94	24.54	66.67	6.48
HRSCD-st.4	13.71	43.97	24.35	65.35	6.78
HBSCD	49.10	175.84	25.25	66.19	7.69
SSCD-l	23.31	189.76	24.69	65.52	7.18
BiSRNet	23.39	189.91	26.95	67.36	9.64
SCanNet	27.90	264.94	25.39	66.71	7.68
MCTNet	26.36	76.71	27.32	67.56	10.06
HGINet	27.70	50.63	27.34	66.28	10.65
CDLNet	12.88	30.10	27.50	67.33	10.43
S^{3}CD	46.27	129.17	29.05	68.86	12.00

TABLE III

EXPERIMENTAL RESULTS COMPARING THE SOTA SUPERVISED METHOD AND S^{3}CD METHOD FOR HI-UCD MINI DATASET

Model	Computation Costs		Accuracy		
	Params(M)	FLOPs(G)	Score(%)	mIoU(%)	Sek(%)
HRSCD-st.3	12.77	42.94	21.10	62.74	3.25
HRSCD-st.4	13.71	43.97	25.62	66.71	8.01
HBSCD	49.10	175.84	29.57	68.30	12.97
SSCD-l	23.31	189.76	25.37	66.12	7.91
BiSRNet	23.39	189.91	25.07	65.24	7.86
SCanNet	27.90	264.94	26.34	65.67	9.49
MCTNet	26.36	76.71	27.21	66.16	10.52
HGINet	27.70	50.63	29.71	67.53	13.50
CDLNet	12.88	30.10	30.47	69.21	13.86
S^{3}CD	46.27	129.17	33.35	70.26	17.53

and SeK. It is noteworthy that S^{3}CD excels in semantic embedding (SeK), significantly surpassing other methods. This can be attributed to the fact that the model learns the semantic information of the same ground objects in different temporal and the change patterns between different ground objects in the pretraining phase, effectively distinguishing between different land-cover types. In particular, for the SECOND dataset, HRSCD-str3 scored the lowest in all three metrics. This is because this method treats CD and semantic segmentation subtasks as independent, overlooking the correlations between the two. And, it is observed that CDLNet ranks second only to S^{3}CD in the mIoU metric. This is because the method integrates temporal change information while acquiring semantic information. It follows that these two factors play a key role in SCD performance, thus affirming the need for joint modeling of these factors in our design of self-supervised frameworks. It is noteworthy that S^{3}CD achieves far better performance than other methods without using complex components in the downstream task. This demonstrates the importance of self-supervised training for SCD.

B. Evolution of Bi-Temporal Semantic Feature in S^{3}CD Pretraining

The advantage of the S^{3}CD self-supervised framework lies in its capacity to enable the model to extract semantic change information from unlabeled bi-temporal images during the pretraining phase. To verify the advantages, this section visualizes the semantic features and semantic changes clustering results across the S^{3}CD pretraining process. Specifically, we employed t-SNE [55] to visualize the feature representations of the backbone output at three time points during the pretraining phase. The visualizations in Figs. 6 and 7 use different colors to represent predicted cluster assignments, enabling qualitative analyses of the model's ability to learn and discriminate semantics change.

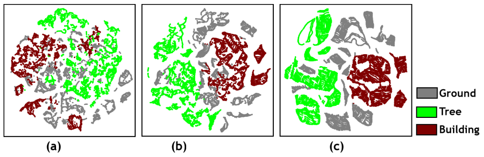


Fig. 6. Visualization of bi-temporal LCLU feature representations during S^{3}CD pretraining on the SECOND dataset with t-SNE. (a) Initial. (b) After Stage 1. (c) Final.

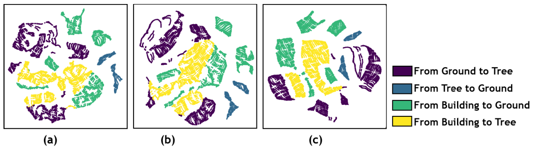


Fig. 7. Visualization of bi-temporal LCLU change patterns representations during S^{3}CD pretraining on the SECOND dataset with t-SNE. (a) Initial. (b) After Stage 1. (c) Final.

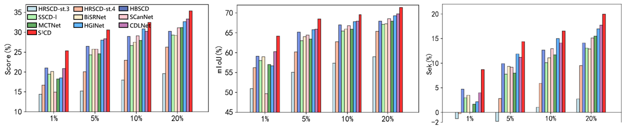


Fig. 8. Results of supervised and S^{3}CD using limited label data on the SECOND dataset.

As illustrated in Fig. 6, initially, features are all mixed and most of the LCLU categories are assigned to a few clusters. As Stage 1 of S^{3}CD pretraining progresses, features scatter more distinctly and cluster assignments become more balanced, suggesting improved category recognition. After the completion of Stage 2 of S^{3}CD pretraining, the model achieved a more compact and well-separated LCLU cluster. This indicates that the S^{3}CD framework effectively enhances the model's ability to distinguish and represent diverse LCLU categories.

Conversely, Fig. 7 reveals a different temporal progression for semantic change pattern representation. After Stage 1 pretraining, the feature representation of semantic change patterns shows minimal differentiation from the untrained initialization. However, after Stage 2 of pretraining, the LCLU change pattern is clearly depicted in the feature representations. This shows the importance of the two-stage training paradigm of S^{3}CD, especially that Stage 2 is crucial for learning robust representations of complex semantic change patterns.

C. Ablation Experiments

To comprehensively evaluate the contributions of individual components in S^{3}CD, we conducted an ablation study on the SECOND dataset involving SCD. In this study, we performed six different experiments to assess the impact of three crucial components. For convenience, we labeled different models as

TABLE IV

RESULTS OF ABLATION EXPERIMENTS ON THE SECOND DATASET

ID	Components				Score(%)	mIoU(%)	Sek(%)
	Stage1	SA	CCP	FD			
1					35.45	71.31	20.08
2	✓				36.22	71.92	20.92
3	✓	✓			37.25	72.44	22.17
4	✓		✓		37.78	72.94	22.70
5	✓	✓	✓		38.25	73.06	23.33
6	✓			✓	36.71	72.19	21.50
7	✓	✓	✓	✓	39.07	73.58	24.28

IDs 1-7, where Model 1 refers to the HRNet network using ImageNet pretrained parameters. All tested models maintained consistency with the experimental parameter settings mentioned earlier. The results of the ablation study are presented in Table IV.

1) *Stage 1 of S^{3}CD (via Instance Loss):* Compared with the baseline model, it suggests that the introduction of the instance loss enhances the capability of the model to extract both change and semantic features.

2) *SA and CCP:* Model 3 and Model 4 are based on Model 2, with the addition of SA and CCP components, respectively. It can be observed that these two components show significant performance improvements in the SCD task. In particular, the CCP module achieves excellent performance improvement, which indicates that learning the change pattern of ground objects enhances the model's ability to discriminate semantic changes. In addition, it can be noted that Model 5, which combines SA and CCP, shows performance improvements that exceed those achieved by either module alone. This synergy suggests that complementary mechanisms for modeling semantic consistency in stable regions (via SA) and capturing semantic change patterns in change regions (via CCP) are both critical for robust SCD.

3) *Feature Distinction:* Model 6 builds upon Model 2 by incorporating an FD mechanism to amplify the distinctions between change and stable features within the bi-temporal feature. Quantitative evaluation showed that FD contributes less to SCD performance compared to other components. However, the comparative analysis between Model 7 and Model 5 indicates that the inclusion of the FD module enhances the performance of SCD. This improvement stems from the synergistic interaction between FD and existing modules, which refines the delineation of changed versus unchanged regions and enhances the semantic learning capabilities of both the SA and CCP components. Furthermore, the role of the FD in improving the model's ability to distinguish changes will be empirically validated in Section VI-D.

D. Effects of Mamba on Semantic Change Pattern Modeling

To validate the effectiveness of using the Mamba network for modeling semantic change patterns, experiments were conducted to analyze the interaction between different modeling mechanisms. Specifically, the following test schemes were applied within the CCP components to model semantic change patterns: Simple concatenation of bi-temporal features, Mamba with TS modeling only, Mamba with TC modeling only, and Mamba with integrated TS and TC modeling. To reduce the impact of other components, the baseline model is set to a pretrained network from Stage 1 of the S^{3}CD framework (i.e., Model 2 in Section V-C). This baseline allows for a direct evaluation of the incremental improvements brought by each modeling strategy, highlighting the advantages of integrating both TC and TS modeling mechanisms for capturing complex semantic change patterns.

TABLE V

COMPARISON OF DIFFERENT APPROACHES FOR MODELING SEMANTIC CHANGE PATTERN ON THE SECOND DATASET

Modeling Mechanism			Score(%)	mIoU(%)	Sek(%)
Concatenation	TS	TC			
			36.22	71.92	20.92
✓			36.82	72.20	21.65
	✓		37.45	72.52	22.42
		✓	37.52	72.64	22.55
	✓	✓	37.78	72.94	22.70

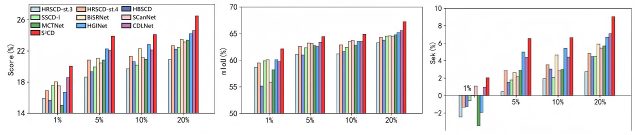


Fig. 9. Results of supervised and S^{3}CD using limited label data on the HRSCD dataset.

In Table V, we list the SCD performance obtained using two modeling mechanisms. Experimental results reveal that while both TC and TS modeling mechanisms demonstrate comparable effectiveness, the TC modeling approach exhibits marginally superior performance. Furthermore, combining both TC and TS modeling mechanisms leads to additional performance improvements, suggesting their complementary roles in capturing different aspects of semantic change patterns.

VI. DISCUSSION

A. Experiments for S^{3}CD Fine-Tuning With Limited Data

The quantity of labeled samples is a crucial factor influencing the performance of deep learning methods. However, in practical applications of SCD, annotated data is often extremely limited. Therefore, this section aims to assess the performance of the proposed S^{3}CD under conditions of limited annotations and compare it with the comparative methods. Experiments were conducted by randomly selecting 1%, 5%, 10%, and 20% labeled samples from the SECOND and HRSCD datasets.

Figs. 8 and 9 present the SCD performance of various methods on the SECOND and HRSCD datasets with different sample ratios. It is evident that as the number of training samples increases, the performance of different models improves to a certain extent. In particular, S^{3}CD consistently maintains the best performance for experiments with different data sizes on both datasets, which proves that the method has better generalization performance and can effectively reduce the dependence on labeled samples. Specifically, in terms of SeK, S^{3}CD outperforms other methods noticeably, which is attributed to the ability of the model to learn to identify

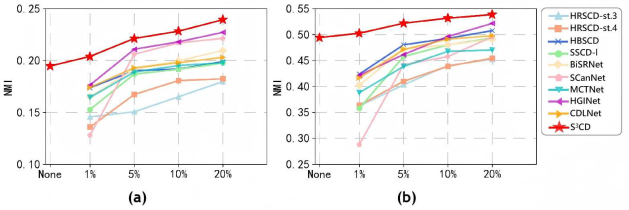


Fig. 10. Comparison of semantic change clustering results between supervised methods with limited label data and S^{3}CD without label fine-tuning on the SECOND dataset. (a) Clustering result of semantic change and (b) clustering result of semantic change of the masking unchanged regions.

semantic changes during pretraining, allowing it to recognize semantic change even under conditions of extreme lack of samples. It is worth noting that, with the aid of the massive unlabeled data, the proposed S^{3}CD framework achieves the equivalent performance of a 20% labeled data comparative method with only 5% training samples. This suggests that S^{3}CD is suitable for applications where tag data is scarce or difficult to obtain.

B. Experiments on Semantic Change Clustering of S^{3}CD

To further explore S^{3}CD's ability to learn semantic change information during the pretraining phase, this section compares the performance of the S^{3}CD (without label fine-tuning) for semantic change clustering with the supervised model described in Section VI-A (trained with limited labeled data).

The semantic change clustering procedure involves concatenating the semantic features extracted from bi-temporal images by the network, followed by the application of k-means clustering to identify distinct semantic change categories. To analyze the clustering efficacy of semantic change patterns, we employ normalized mutual information (NMI) [56], a widely adopted clustering metric in unsupervised learning evaluation. NMI measures the statistical dependence between the predicted cluster assignments and ground-truth semantic change labels, with values ranging from 0 to 1, where higher scores indicate superior clustering performance and more accurate semantic change pattern detection. As shown in Fig. 10(a), quantitative results show that S^{3}CD pretraining significantly improves the network's ability to detect semantic changes. Notably, when semantic change annotations are limited to less than 10%, S^{3}CD without requiring label fine-tuning performs better than most supervised methods in terms of semantic change clustering.

It is noteworthy that during the semantic change clustering process, all unchanged regions are grouped into a single cluster. This methodological decision introduces evaluation complexities since unchanged regions encompass diverse subtypes: natural changes (phenological changes), state changes without semantic changes (e.g., architectural renovations), and regions exhibiting no change whatsoever. These nuanced distinctions within the "unchanged" category may influence clustering performance metrics. In order to further explore the semantic clustering ability of S^{3}CD, we conducted additional experiments masking the unchanged regions, focusing only on differential semantic change patterns. As shown in Fig. 10(b), S^{3}CD without label fine-tuning exhibits comparable performance to supervised methods with limited samples (sample sizes below 20%). These findings underscore the effectiveness of S^{3}CD in robust SCD and highlight its practical significance in real-world scenarios where semantic change annotations are scarce or completely unavailable.

TABLE VI

Comparative Experimental Results of BCD on the SECOND Dataset

	Model	Label	F1(%)	Kappa(%)
Supervised	FC-EF	✓	41.71	31.24
	SiamUnet_diff	✓	41.93	32.15
	SiamUnet_conc	✓	42.59	33.45
Unsupervised	PCA-k-mean		29.20	11.86
	DCVA		24.33	13.57
	S^{3}CD-unsupervised		43.71	34.10

TABLE VII

Comparative Experimental of BCD Obtained by S^{3}CD With and Without FD on the SECOND Dataset

Model	F1(%)	Kappa(%)
S^{3}CD-unsupervised (without FD)	40.62	31.21
S^{3}CD-unsupervised	43.71	34.10

C. Evaluation on BCD

Self-supervised learning enables the model to learn the ability to distinguish between change and no change from unlabeled data. Using only self-supervised pretrained parameters without fine-tuning, S^{3}CD can function as a BCD-oriented self-supervised CD algorithm to determine whether land-cover has changed. To evaluate S^{3}CD's ability to locate changes, we conducted experiments on the SECOND dataset and refer to this variant as S^{3}CD-unsupervised. The self-supervised pretraining setup remains consistent with Section V-A, with no annotated fine-tuning. Specifically, a pretrained feature extractor is used to extract depth features from bi-temporal images, and the Otsu thresholding method [57] is applied to generate the final binary change map. To verify the superiority of the proposed method, comparative experiments were conducted on five other popular BCD methods, including FC-EF [58], FC-Siam-Conc [58], FC-Siam-Diff [58], PCA-k-means [59], and DCVA [60]. During the supervised training period, data augmentation strategies were applied, including rotation, cropping, and color enhancement. The evaluation was based on two metrics: $F1$-score ($F1$) and Kappa coefficient (Kappa), with detailed definitions available in [60].

As shown in Table VI, as expected, supervised methods outperform common unsupervised methods in most metrics. It is worth noting that in this dataset, the S^{3}CD-unsupervised model is the best performer among all methods, even if compared to the supervised methods.

D. Effects of the FD on BCD

Ablation experiments (see Section V-C) demonstrate that the FD enhances the performance of SA and CCP components by improving the network's ability to recognize changes, thereby further improving the network's overall performance in SCD. To more intuitively illustrate the impact of the FD

TABLE VIII
COMPARATIVE EXPERIMENTAL RESULTS OF SELF-SUPERVISED MODELS IN THE SECOND DATASET

Model	SCD			BCD	
	Score	mIoU	SeK	F1	Kappa
Base (ImageNet)	34.86	70.79	19.46	48.42	38.07
SimSiam	35.22	71.16	19.81	51.08	41.46
SimCLR	35.32	71.11	19.99	51.37	41.84
BYOL	35.36	71.19	20.01	51.74	42.13
MoCo v2	35.40	70.94	20.11	50.52	40.94
TWIST	35.20	71.00	19.86	50.33	40.58
Contrastive Clustering	35.11	70.65	19.88	47.63	38.77
DenseCL	35.51	71.15	20.23	51.02	41.66
GLCL	36.02	71.17	20.95	51.94	42.48
SACNet	36.41	71.52	21.36	52.91	43.40
SSLchange	36.02	71.17	20.95	51.94	42.48
GRESS	36.07	71.21	21.02	50.63	39.21
S^{3}CD-Res	37.82	72.75	22.85	55.25	46.40
S^{3}CD-HR	39.07	73.58	24.28	58.72	50.09
GLCL-unsupervised	-	-	-	33.28	18.23
SACNet-unsupervised	-	-	-	33.43	22.80
GRESS-unsupervised	-	-	-	27.52	15.21
S^{3}CD-Res-unsupervised	-	-	-	40.52	30.36
S^{3}CD-HR-unsupervised	-	-	-	43.71	34.10

on the ability of the network to detect changes, we conducted additional ablation experiments on the SECOND dataset in this section. Specifically, we removed the FD component and evaluated the network's performance on the BCD task. The experimental setup is consistent with that described in Section V-A. Table VII shows that removing the FD lowers all performance metrics for detecting change location. This highlights FD's crucial role in accurately distinguishing between change and unchanged regions, confirming its importance in the overall architecture.

E. Comparative Experiments of Self-Supervised Methods

In order to fully analyze the performance of the proposed method, this section compares S^{3}CD with other SOTA self-supervised methods (as seen in Table VII), including SimSiam [34], BYOL [35], SimCLR [44], MoCo v2 [61], TWIST [47], Contrastive Clustering [46], DenseCL [62], SSLChange [63], GLCL [64], SACNet [65], and GRESS [66]. Among them, the first eight methods only pretrain the encoder (i.e., its encoded feature size is 1/64 of the input image, which is difficult to use directly for pixel-level interpretation tasks), while GLCL, GRESS, and SACNet can pretrain the entire encoder–decoder structure. For fair comparison, ResNet50 [67] is used as the encoder for all self-supervised algorithms, and UperNet [68] is employed as the decoder. And the HRNet in S^{3}CD is correspondingly replaced, denoted as S^{3}CD-Res, while the unreplaced version is referred to as S^{3}CD-HR. In addition, in Table VII, "Base" refers to the initialization parameters of the encoder that are pretrained on ImageNet.

The aforementioned comparative algorithms independently input images from different temporal frames during pretraining to acquire semantic representations and adopt the same optimization objectives during downstream fine-tuning. Specifically, for the supervised fine-tuning phase of the SCD task, the amount of downstream labeled samples is kept the same as in Section V-A. While for the BCD task, we use 1% of the labeled samples for fine-tuning, and the parameter settings are kept consistent with the BCD branch in the supervised SCD framework. In addition, the performance of GLCL, SACNet, GRESS, S^{3}CD-Res, and S^{3}CD-HR (all of which can be pretrained to obtain high-resolution pixel-level feature representations) is tested under unsupervised conditions for the BCD task, denoted by the suffix "-unsupervised" in Table VIII.

From Table VIII, it can be seen that S^{3}CD-Res has all the metrics higher than other self-supervised methods in both tasks, making it more suitable for CD. It is worth mentioning that, with the exception of GLCL, SACNet, GRESS, and S^{3}CD, none of the generic self-supervised methods (pretrained encoders only) in the unlabeled case can be directly applied to the task of CD on the SECOND dataset. These results highlight the robustness and semantic modeling capabilities of S^{3}CD-Res, and provide an efficient and reliable solution for self-supervised SCD tasks.

VII. CONCLUSION

Self-supervised learning methods have demonstrated remarkable potential across various domains in remote sensing. However, their application in the field of SCD has not been fully explored. Therefore, this article is devoted to the introduction of self-supervised methods into the SCD task to learn generalized temporal-invariant features from a large number of unlabeled images in order to reduce the dependence on labeled samples, thus further improving the recognition accuracy of the deep learning model as well as increasing the utility of the model in a small number of label situations. In particular, to overcome the limitations of self-supervised methods that are restricted to acquiring scene-level information, S^{3}CD employs a two-stage process (Stage 1: learning coarse semantic features from a single-temporal image and Stage 2: capturing fine-grained semantic change information from bi-temporal images) to guide the model in learning Semantic change information at the scene, pixel, and prototype levels. Experiments on three benchmark SCD datasets confirm that our method achieves superior performance compared to SOTA alternatives. Importantly, in the task of limited labeled data, our approach exhibits significant advantages over other SCD models, suggesting that S^{3}CD may have great practical application. Furthermore, in the BCD task, our method (without label) outperforms both supervised and unsupervised methods on several metrics, closing the gap between unsupervised and supervised methods.

Frankly, our approach still has some limitations. For instance, we did not design a sophisticated fine-tuning framework for SCD, which could limit the performance of the pretrained model. Therefore, in the future, we will focus on developing more advanced frameworks for SCD and concurrently constructing datasets suitable for self-supervised SCD tasks.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for the insightful suggestions, which significantly improved the quality of this article.

References

[1] F. Bovolo and L. Bruzzone, "The time variable in data fusion: A change detection perspective," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 8–26, Sep. 2015.

[2] E. Khankeshizadeh, A. Mohammadzadeh, A. Moghimi, and A. Mohsenifar, "FCD-R2U-Net: Forest change detection in bi-temporal satellite images using the recurrent residual-based U-Net," *Earth Sci. Informat.*, vol. 15, no. 4, pp. 2335–2347, Dec. 2022.

[3] X. Huang, J. Li, J. Yang, Z. Zhang, D. Li, and X. Liu, "30 M global impervious surface area dynamics and urban expansion pattern observed by Landsat satellites: From 1972 to 2019," *Sci. China Earth Sci.*, vol. 64, no. 11, pp. 1922–1933, Nov. 2021.

[4] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[5] X. Lu, Y. Yuan, and X. Zheng, "Joint dictionary learning for multispectral change detection," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 884–897, Apr. 2017.

[6] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12084–12098, Nov. 2022.

[7] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understand.*, vol. 187, Oct. 2019, Art. no. 102783.

[8] K. Yang et al., "Asymmetric Siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609818.

[9] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, "SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102465.

[10] Q. Wang, W. Huang, X. Zhang, and X. Li, "GLCM: Global–local captioning model for remote sensing image captioning," *IEEE Trans. Cybern.*, vol. 53, no. 11, pp. 6910–6922, Nov. 2023.

[11] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, no. 6, pp. 1082–1086, Nov. 2019.

[12] S. Yuan et al., "Relational part-aware learning for complex composite object detection in high-resolution remote sensing images," *IEEE Trans. Cybern.*, vol. 54, no. 10, pp. 6118–6131, Oct. 2024.

[13] C. Zhang, K.-M. Lam, T. Liu, Y.-L. Chan, and Q. Wang, "Structured adversarial self-supervised learning for robust object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5613720.

[14] C. Zhang, T. Liu, J. Xiao, K.-M. Lam, and Q. Wang, "Boosting object detectors via strong-classification weak-localization pretraining in remote sensing imagery," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–20, 2023.

[15] Z. Huang, J. Shi, and X. Li, "Quantum few-shot image classification," *IEEE Trans. Cybern.*, vol. 55, no. 1, pp. 194–206, Jan. 2025.

[16] Y. Xie, Z. Xu, J. Zhang, Z. Wang, and S. Ji, "Self-supervised learning of graph neural networks: A unified review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2412–2429, Feb. 2023.

[17] L. Wan, Y. Xiang, W. Kang, and L. Ma, "A self-supervised learning pretraining framework for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5630116.

[18] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-D-Swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411415.

[19] Y. Chen and L. Bruzzone, "Self-supervised change detection in multi-view remote sensing images," 2021, *arXiv:2103.05969*.

[20] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[21] I. Katircioglu, H. Rhodin, V. Constantin, J. Spörri, M. Salzmann, and P. Fua, "Self-supervised human detection and segmentation via background inpainting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9574–9588, Dec. 2022.

[22] S. Sun, T. Wang, H. Yang, and F. Chu, "An environmentally adaptive and contrastive representation learning method for condition monitoring of industrial assets," *IEEE Trans. Cybern.*, vol. 54, no. 3, pp. 1484–1496, Mar. 2024.

[23] Y. Qu, J. Li, X. Huang, and D. Wen, "TD-SSCD: A novel network by fusing temporal and differential information for self-supervised remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5407015.

[24] J. Yang and X. Huang, "30 M annual land cover and its dynamics in China from 1990 to 20," *Earth Syst. Sci. Data Discuss.*, vol. 2021, pp. 1–29, Apr. 2021.

[25] E. Veldkamp, M. Schmidt, J. S. Powers, and M. D. Corre, "Deforestation and reforestation impacts on soils in the tropics," *Nature Rev. Earth Environ.*, vol. 1, no. 11, pp. 590–605, Sep. 2020.

[26] L. Bruzzone, R. Cossu, and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognit. Lett.*, vol. 25, no. 13, pp. 1491–1500, Oct. 2004.

[27] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral–spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[28] Q. Zhu et al., "Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022.

[29] O. Ahlqvist, "Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: A study of 1992 and 2001 U.S. National land cover database changes," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 1226–1241, Mar. 2008.

[30] H. Xia, Y. Tian, L. Zhang, and S. Li, "A deep Siamese postclassification fusion network for semantic change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622716.

[31] S. Xiang, M. Wang, X. Jiang, G. Xie, Z. Zhang, and P. Tang, "Dual-task semantic change detection for remote sensing images using the generative change field module," *Remote Sens.*, vol. 13, no. 16, p. 3336, Aug. 2021.

[32] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620014.

[33] B. Xiao, J. Hu, W. Li, C.-M. Pun, and X. Bi, "CTNet: Contrastive transformer network for polyp segmentation," *IEEE Trans. Cybern.*, vol. 54, no. 9, pp. 5040–5053, Sep. 2024.

[34] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15750–15758.

[35] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[36] J. Gui et al., "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9052–9071, Dec. 2024.

[37] Y. Chen and L. Bruzzone, "Self-supervised change detection by fusing SAR and optical multi-temporal images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3101–3104.

[38] S. Saha, P. Ebel, and X. Xiang Zhu, "Self-supervised multisensor change detection," 2021, *arXiv:2103.05102*.

[39] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," 2021, *arXiv:2106.10605*.

[40] Y. Chen and L. Bruzzone, "A self-supervised approach to pixel-level change detection in bi-temporal RS images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4413911.

[41] M. Zhao, X. Hu, L. Zhang, Q. Meng, Y. Chen, and L. Bruzzone, "Beyond pixel-level annotation: Exploring self-supervised learning for change detection with image-level supervision," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5614916.

[42] C. Wang, S. Du, W. Sun, and D. Fan, "Self-supervised learning for high-resolution remote sensing images change detection with variational information bottleneck," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5849–5866, 2023.

[43] J. Wang et al., "MaCon: A generic self-supervised framework for unsupervised multimodal change detection," *IEEE Trans. Image Process.*, vol. 34, pp. 1485–1500, 2025.

[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[45] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5693–5703.

[46] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, “Contrastive clustering,” in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 10, pp. 8547–8555.

[47] F. Wang, T. Kong, R. Zhang, H. Liu, and H. Li, “Self-supervised learning by estimating twin class distribution,” *IEEE Trans. Image Process.*, vol. 32, pp. 2228–2236, 2023.

[48] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” 2023, *arXiv:2312.00752*.

[49] S. Tian, Y. Zhong, Z. Zheng, A. Ma, X. Tan, and L. Zhang, “Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application,” *ISPRS J. Photogramm. Remote Sens.*, vol. 193, pp. 164–186, Nov. 2022.

[50] Z. Jiang, B. Wang, P. Zhang, Y. Wu, Z. Ye, and H. Yang, “Semantic enhancement and change consistency network for semantic change detection in remote sensing images,” *Int. J. Digit. Earth*, vol. 18, no. 1, Aug. 2025, Art. no. 2496790.

[51] L. Ding, J. Zhang, H. Guo, K. Zhang, B. Liu, and L. Bruzzone, “Joint spatio-temporal modeling for semantic change detection in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610814.

[52] W. Liu, Z. Kang, J. Liu, Y. Lin, Y. Yu, and J. Li, “A multitask CNN-transformer network for semantic change detection from bitemporal remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5647215.

[53] J. Long, M. Li, X. Wang, and A. Stein, “Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images,” *ISPRS J. Photogramm. Remote Sens.*, vol. 211, pp. 318–335, May 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271624001709

[54] J. Zhang et al., “Joint content-aware and difference-transform lightweight network for remote sensing images semantic change detection,” *Inf. Fusion*, vol. 123, Nov. 2025, Art. no. 103276.

[55] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[56] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.

[57] N. Otsu, “A threshold selection method from gray-level histograms,” *Automatica*, vol. 11, nos. 285–296, pp. 23–27, 1975.

[58] R. C. Daudt, B. Le Saux, and A. Boulch, “Fully convolutional Siamese networks for change detection,” in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[59] L. Bruzzone and D. F. Prieto, “Automatic analysis of the difference image for unsupervised change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[60] S. Saha, F. Bovolo, and L. Bruzzone, “Unsupervised deep change vector analysis for multiple-change detection in VHR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[61] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” 2020, *arXiv:2003.04297*.

[62] X. Wang, R. Zhang, C. Shen, and T. Kong, “DenseCL: A simple framework for self-supervised dense visual pre-training,” *Vis. Informat.*, vol. 7, no. 1, pp. 30–40, Mar. 2023.

[63] Y. Zhao, T. Celik, N. Liu, F. Gao, and H.-C. Li, “SSLChange: A self-supervised change detection framework based on domain adaptation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5647814.

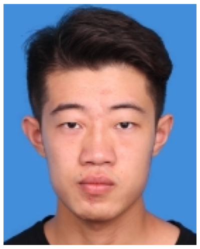
[64] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.

[65] D. Chen et al., “SACNet: A novel self-supervised learning method for shadow detection from high-resolution remote sensing images,” *J. Geovis. Spatial Anal.*, vol. 9, no. 1, p. 14, Jun. 2025.

[66] Y. Chen, W. Wu, L. Ou-Yang, R. Wang, and S. Kwong, “GRESS: Grouping belief-based deep contrastive subspace clustering,” *IEEE Trans. Cybern.*, vol. 55, no. 1, pp. 148–160, Jan. 2025.

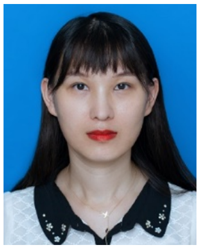
[67] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[68] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.



Yang Qu received the B.S. degree in Earth information science and technology from Henan Polytechnic University, Jiaozuo, Henan, China, in 2018. He is currently pursuing the Ph.D. degree in remote sensing with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include remote sensing image processing, change detection, and deep learning.



Jiayi Li (Senior Member, IEEE) received the B.S. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University. She has authored more than 60 peer-reviewed articles (science citation index (SCI) articles) in international journals. Her research interests include hyperspectral imagery, sparse representation, computational vision and pattern recognition, and remote sensing images.

Dr. Li is a reviewer for more than 30 international journals, including IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Image Processing, IEEE Transactions on Cybernetics, RSE, and ISPRS-J. She is the Young Editorial Board Member of *Geo-Spatial Information Science* (GSIS) and a Guest Editor of *Remote Sensing* (an open access journal from MDPI) and *Sustainability* (an open access journal from MDPI).

Xiaofeng Pan received the Ph.D. degree in environmental science from the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China, in 2010.

He is currently with Shenzhen Ecological and Environmental Monitoring Center Station of Guangdong Province, Shenzhen, China, mainly engaged in monitoring the quality of the ecological environment and comprehensive analysis of data.

Xin Huang (Fellow, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009.

He is currently a Full Professor at Wuhan University, where he teaches remote sensing and image interpretation. He is the Head of the Institute of Remote Sensing Information Processing (IRSIP), Wuhan University. He has published more than 230 peer-reviewed articles (SCI articles) in international journals. He was identified as the “Most Cited Chinese Researchers” by Elsevier and the “Highly Cited Researchers” by Clarivate. He produced China Land-Cover Data (CLCD) from 1980 to 2023, an open-source 30-m annual land-cover dataset of China, and the global impervious surface area from 1972 to 2023 (GISA) with the spatial resolution of 30 m. He proposed a series of 3-D reconstruction and interpretation methods for the multiview satellite images, which have been used to generate the 3-D high-resolution urban land-cover maps (ULCM) over China’s 50 major cities. His research interests include remote sensing image processing methods and applications.

Dr. Huang was elected as an IEEE Fellow, for his contributions to “machine learning for remote sensing.” He won the First Prize of China Remote Sensing Outstanding Achievement Award, the First Prize of China Geographical Information Science and Technology Progress Award, and the First Prize for China Surveying and Mapping Science and Technology Progress Award. He was a recipient of the Boeing Award from American Society for Photogrammetry and Remote Sensing (ASPRS) in 2010 and the John I. Davidson President’s Award from ASPRS in 2018. He was the winner of the IEEE GRSS Data Fusion Contest in 2014 and 2021. He was an Associate Editor of IEEE Geoscience and Remote Sensing Letters (2014–2020) and IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2018–2022) and now serves as an Associate Editor for IEEE Transactions on Geoscience and Remote Sensing (since 2022). He is an Editorial Board Member of *Remote Sensing of Environment* (since 2019).