TD-SSCD: A Novel Network by Fusing Temporal and Differential Information for Self-Supervised Remote Sensing Image Change Detection

Yang Qu, Jiayi Li, *Senior Member, IEEE*, Xin Huang, *Senior Member, IEEE*, and Dawei Wen

***Abstract*— The change detection of remote sensing images has a wide range of applications in many fields. In recent years, deep learning has become one of the most powerful tools for remote sensing change detection due to its excellent feature learning capability. However, most deep learning methods require a lot of labeled data for the training, which is time-consuming and labor-intensive. Recently, a new learning paradigm—self-supervised learning—has become one of the hot topics in the field of change detection due to its ability to learn feature representations by training with a large amount of unlabeled data and without a large number of sample annotations. However, the existing methods for self-supervised learning are usually designed for natural image processing and are less considered for change detection in more complex scenes (e.g., remote sensing imagery). Therefore, in this article, we propose a novel network by fusing temporal and differential information for self-supervised contrastive learning change detection, namely, TD-SSCD. Specifically, TD-SSCD aims to mine information from the bitemporal images and their differential images (DIs) in a self-supervised learning framework, and it gradually learns the potential correlations between them through an alternating iteration learning strategy. The experimental results based on the Onera Satellite Change Detection (OSCD) and Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI) datasets show that the proposed method outperforms the current state-of-the-art (SOTA) unsupervised and self-supervised change detection (SSCD) methods. Benefiting from pretraining on unlabeled samples, the method closes the gap between unsupervised and supervised change detection.**

***Index Terms*— Contrastive learning, remote sensing image change detection, self-supervised learning, temporal and differential.**

I. Introduction

CHANGE detection has become one of the most important topics in the field of remote sensing [1], [2]. It is the technology of observing and identifying the information about land-surface changes by analyzing two (or more) images that are obtained at different observation times in the same area. Change detection in remote sensing images has been widely applied in various applications, such as land management [3], environmental monitoring [4], disaster assessment [5], and urban planning [6]. The existing remote sensing image change detection techniques can be roughly divided into supervised and unsupervised ones based on whether labeled training samples are used or not [7]. The supervised techniques have achieved satisfactory results in change detection [8]. However, supervised techniques require considerable prior knowledge [9]. In other words, they require more training samples for learning the features of remote sensing images. In practical applications, it is difficult to collect a large number of labeled samples of remote sensing images, which often limits the practicability of the supervised methods. Thus, unsupervised remote sensing change detection techniques that do not depend on the availability of labeled samples are receiving increased attention.

The traditional unsupervised remote sensing change detection methods tend to use clustering or a threshold to determine the changed or unchanged regions [10]. Unfortunately, most of these methods are designed with the individual pixel as the elementary unit and rely on hand-crafted features, resulting in poor robustness in complex scenes [11]. In recent years, deep learning techniques have been widely used in remote sensing due to their powerful feature learning capabilities [12], [13], [14], [15], and many unsupervised change detection methods based on deep learning have been developed and have achieved impressive results. A pretrained convolutional neural network (CNN) was applied to extract spectral–spatial features from bitemporal images, and then, the conventional change vector analysis (CVA) was employed to identify changes [16]. In addition to these various CNNs, autoencoders [17] and generative adversarial networks (GANs) [18] are widely used in unsupervised change detection tasks [19]. For example, Gong et al. [20] transformed heterogeneous images into a shared-latent space via variational autoencoders and then used GAN to obtain more accurate change maps. However, some studies have shown that the commonly used unsupervised change detection methods, such as GAN [21], focus more on the pixels themselves rather than abstract deep feature representations [22]. To address this problem, researchers have proposed change detection methods based on transfer learning

Manuscript received 10 August 2022; revised 9 February 2023 and 24 July 2023; accepted 25 September 2023. Date of publication 27 September 2023; date of current version 10 October 2023. The work was supported in part by the National Natural Science Foundation of China under Grant 42071311, Grant 41971295, and Grant 42271328; in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100031; and in part by Wuhan 2022 Shuguang Project under Grant 2022010801020123. *(Corresponding author: Jiayi Li.)*

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

Jiayi Li is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: zjjercia@whu.edu.cn).

Dawei Wen is with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430070, China (e-mail: daweiwen_whu@163.com).

Digital Object Identifier 10.1109/TGRS.2023.3319961

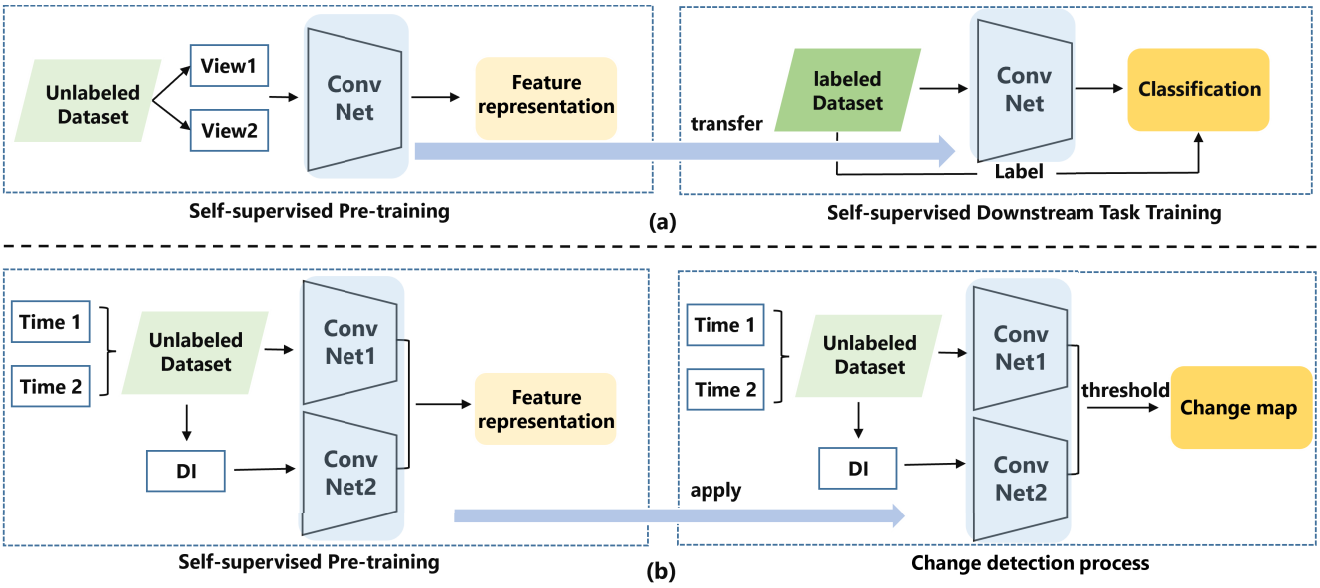

Fig. 1. Self-supervised learning paradigm. (a) Process of self-supervised learning. (b) Proposed TD-SSCD change detection method DI.

to extract the deep features of remote sensing images by using a pretrained model to obtain the change information [16]. However, these methods based on transfer learning greatly rely on the transferability of the pretrained model.

More recently, self-supervised learning method provides a new paradigm, as shown in Fig. 1(a). First, self-supervised model learns universal feature representations from a massive unlabeled image data and then transfers to downstream tasks to achieve similar (even better) performance to supervised learning on downstream tasks with a limited number of labeled samples. The way to construct sample pairs is one of the key points of self-supervised learning [23]. For example, Chen et al. [24] proposed MoCov3, which encourages features from different views of the same image (positive pairs) to be closer and feature representations from different images (negative pairs) to be more exclusive, to guide the model to learn semantical invariant features of image. There are also self-supervised methods that do not require negative pairs, e.g., SimSiam [25] and bootstrap your own latent (BYOL) [26], which also achieve good model performance by simply encouraging the maximization, the similarity between positive pairs, significantly reducing the computational cost. Recently, self-supervised learning methods have been applied to remote sensing change detection tasks [27], [28]. Specifically, the existing self-supervised learning-based change detection methods usually follow the self-supervised pretraining process used in scene classification to obtain a well-trained feature extractor. Chen and Bruzzone [28] used the BYOL method for change detection, showing similar performance to the supervised method. It is worth noting that the above self-supervised change detection (SSCD) methods work under the assumption that the proportion of change regions in the image is small. By self-supervised learning of temporal invariant features (i.e., semantic features of the ground objects), the semantic features of the bitemporal remote sensing images are subtracted during reasoning to determine whether the objects have changed. In this process, the features in the differential space of the bitemporal images are ignored. From this point of view, the SSCD methods might obtain better results if the information from the bitemporal images can be fully utilized. It has been shown in several studies that change detection with early fusion strategies [using differential images (DIs) or connected bitemporal images] is effective in retaining the change information from the original images, but, unfortunately, they may retain some pseudochanges at the same time [29]. In summary, the change detection methods with early and late feature fusion strategies have their own respective advantages and disadvantages, and their complementation is expected to achieve better results [30]. Thus, effectively combining the temporal images and their differential features may be a promising solution for improving the performance of SSCD methods. However, this point has not been considered in the existing unsupervised change detection literature.

In view of this, in this article, we propose a novel SSCD scheme by fusing the temporal and differential feature extraction branches, called TD-SSCD. As illustrated in Fig. 1(b), TD-SSCD learns the temporal change and differential features of the bitemporal images in the pretraining phase, then uses the two well-trained networks to calculate the change probability, and obtains the final binary change map based on an appropriate threshold. Among them, the pretraining process is divided into two stages. In the first stage, we make the model learn the temporal invariance feature of bitemporal images by self-supervised learning. This stage focuses on mining the features of temporal images themselves. Then, in the second stage, we introduce DIs to the self-supervised learning of temporal information in the first stage to realize the joint learning of temporal and differential information in the self-supervised framework. It is worth mentioning that self-supervised frameworks usually use data enhancement techniques (such as cropping, color distortion, or random noise) to obtain multiple views from the same image and then learn the correlation between different views. In contrast, in the method proposed in this article, we propose a new multiple views

construction scheme in order to achieve joint SSCD of temporal and its differential information. We consider that the feature mappings of early fusion (temporal feature learning) and late fusion (differential feature learning) can represent change information from different perspectives and can be considered as multiple views of SSCD. Furthermore, in order to better accommodate the correlation between temporal-differential multiple views during network training, we propose a new training strategy: 1) propose a distribution consistency loss to learn the feature representation; 2) design an update loss and combine it with the distribution consistency loss to form a balanced recurrent architecture to guide the two feature extraction modules to learn from each other during the training process to mine potential relationships between temporal and DIs; and 3) introduce a distribution sharpness component to dynamically suppress error messages and highlight areas that change during training.

The main contributions of this article are summarized as follows.

1) We propose a new SSCD scheme, namely, TD-SSCD. Unlike the existing SSCD methods, this method uses temporal images and their DIs as the self-supervised learning multiple views, aiming to fully mine the change features of temporal images. The method uses two different self-supervised frameworks in pretraining to learn and fuse temporal and differential information, respectively. This self-supervised pretraining process provides a new idea for self-supervised remote sensing change detection.
2) In order that the TD-SSCD network can better utilize the change information in the temporal and DIs, we propose a recurrent alternating self-supervised training method. It uses bitemporal and DIs as positive samples and guides the network to learn and fuse the potential correlations of temporal and DIs through a combination of cyclic alternating loss functions. Meanwhile, we propose a distribution sharpness component to improve the reliability of the change feature distribution.

The rest of this article is organized as follows. Section II presents the related works in change detection and self-supervised learning. Section III introduces the proposed TD-SSCD change detection method in detail. The experimental results obtained on two different datasets and the related comparisons with supervised and unsupervised methods are provided in Section IV. Finally, in Section V, we draw our conclusions.

II. Related Works

In this section, we introduce the methods related to unsupervised change detection and the self-supervised learning.

A. Unsupervised Change Detection

The traditional change detection algorithms commonly use the approach of pixelwise image differencing and adopt clustering or threshold segmentation mechanisms to identify the changed and unchanged areas [31]. Among these methods, a popular one is to combine principal component analysis (PCA) with k-means (i.e., PCA-k-means) to remove noise and the unimportant information in the images by transforming the bitemporal images through PCA and then classifying them with k-means [32]. However, most of these methods compare individual pixels, without considering the relevant information between pixels in the neighborhood. As a result, the resulting change map often contains a large amount of "salt-and-pepper" noise. Thonfeld et al. [33] developed a robust CVA (RCVA) method by considering the surrounding neighborhood of each pixel. However, this method cannot achieve effective feature representation in complex scenes. To obtain a robust feature representation, Bovolo [34] proposed an object-based parcel CVA (PCVA) method, which performs independent hierarchical segmentation of multitemporal images to utilize the spectral and spatial information by encoding the spatial context of pixels. Although all these methods emphasize the importance of spatial context information and neighborhood information, they are based on manually designed features, which are usually not robust.

In recent years, deep learning has been widely applied in remote sensing change detection tasks due to its excellent performance in feature capture and expression [35], [36], [37]. However, deep learning methods usually require a large number of training samples. To solve this problem, some studies have used a preclassification approach to obtain a coarse change map and then select high-confidence samples from the change map to train the network to obtain the final change classification map. For example, Li et al. [37] proposed an unsupervised change detection method based on deep learning, which uses wavelet features to identify changed and unchanged pixels, and then trains the network with patches centered on these pixels as samples. Gao et al. [38] combined a traditional method with deep learning and used the CVA to identify samples with high confidence, and then extracted the pixel features with a high likelihood of change by the deep learning technique, and used the slow feature analysis to highlight the change feature components during the training process. These methods can effectively capture the feature of an image and obtain better results than the traditional ones. However, the quality of the change maps in these deep learning change detection methods is affected by the preclassification change maps. Unfortunately, in complex regions, the results obtained by the traditional methods are not always reliable. To solve this problem, some change detection methods based on transfer learning have been proposed. For example, Li et al. [17] proposed the deep CVA (DCVA) method, which extracts the deep features of the target scene through the deep learning model trained by other tasks, and then compares and analyzes the bitemporal deep change features to obtain the final classification map. However, it seems difficult to obtain satisfactory results when the pretrained dataset and the target task differ significantly. To alleviate this problem, Du et al. [39] proposed a two-stage transfer learning change detection strategy, in which the pretrained model is fine-tuned on the target dataset before the change detection. However, the problem of model performance limitation caused by the difference between the source domain and target domain cannot be solved completely.

B. Self-Supervised Learning

In recent years, self-supervised learning has made much progress, with encouraging results in multiple computer vision tasks [41], [42]. Self-supervised learning allows us to pretrain using the large number of available unlabeled images. Among the self-supervised learning methods, contrastive learning has been successfully applied to the field of natural image processing. As shown in Fig. 1(a), self-supervised contrastive learning consists of two steps: 1) pretraining and 2) fine-tuning for the downstream tasks. Firstly, a self-supervised framework is designed to learn the feature representation from unlabeled image data. Specifically, this allows the network to learn the invariance of the transformation by forcing the different views of the same image (positive sample pairs) to be similar and the different images (negative sample pairs) to be dissimilar. In addition, in this step, the model is able to capture low-level and high-level features that are useful for other downstream tasks. Afterward, the pretrained model can be transferred to the downstream task, thus achieving a performance that is similar to that of supervised learning in the downstream task, with limited labeled samples.

Inspired by this, several studies have applied this self-supervised paradigm to various tasks in remote sensing, and obtained promising results with only a small number of labeled samples. For example, Tao et al. [43] used a self-supervised framework for remote sensing scene classification by using a large unlabeled dataset in the pretraining, and then fine-tuned the well-trained model using a small number of labeled samples to transfer the learned representation to the target task. In the finite labeled data task, this method obtained better results than the traditional supervised methods. Li et al. [44] proposed a global style and local matching contrastive learning network (GLCNet), which guides the model pretraining process through global style and local matching modules to learn multiscale features of the remote sensing images. The GLCNet has shown its superiority over some supervised learning and self-supervised methods on semantic segmentation. Notably, these studies employed data augmentation schemes, such as random noise or color distortion, to simulate the temporal changes in remote sensing images, in order to facilitate the model to learn the temporal invariance of remote sensing images during the pretext process. However, the temporal differences in remotely sensed images are mainly the texture and color differences caused by the complicated imaging conditions, which cannot be truly simulated by the conventional image transformation approaches. Fortunately, remote sensing change detection tasks can naturally provide multiple views (i.e., bitemporal remote sensing images collected at different times in the same region) and do not require data transformation. Therefore, some change detection studies have directly used pre- and postchange images as the multiple views to guide the model to learn the temporal invariance of remote sensing images. For instance, Chen and Bruzzone [28] proposed an SSCD method based on multiple unlabeled views, which achieved a better result than the state-of-the-art (SOTA) unsupervised approach by using the BYOL [26] framework. Saha et al. [27] proposed a self-supervised learning change detection method that combines deep clustering, a Siamese network, and contrastive learning strategies to obtain a satisfactory performance with only a small amount of unlabeled data.

The experimental results reported in the above studies have preliminarily indicated that change detection based on self-supervision has great potential. However, there are still some problems with the existing SSCD methods. To be specific, these SSCD methods are similar to the unsupervised change detection methods based on transfer learning, i.e., a well-trained convolutional network (obtained by pretraining) is used to extract and compare the deep features of bitemporal images (i.e., late fusion). However, although the convolution operation has a strong nonlinear fitting capability, it may lose part of the original image information and lead to some change information being ignored [46]. Since there have been few studies on SSCD, this problem has not been investigated in the self-supervised framework. However, on the other hand, this issue has been addressed in the supervised methods. For example, Daudt et al. [47] designed a late fusion change detection method, which extracts the features of bitemporal images through a depthwise separable fully convolutional network and then fuses the features to generate a change map. Similarly, Gadzicki et al. [48] proposed three different phase fusion networks [i.e., fully convolutional early fusion (FC-EF), fully convolutional Siam-conc (FC-Siam-conc), and fully convolutional Siam-difference (FC-Siam-diff)] to explore the advantages of different feature fusion schemes in change detection tasks. The results showed that the late fusion-based methods can obtain good results. However, in some scenarios, the early fusion approach can obtain a better performance [49]. This means that different fusion schemes have their respective advantages and disadvantages in different scenarios. In this article, inspired by this observation, we propose a new self-supervised learning framework by simultaneously considering the bitemporal images and their differential features to fully explore the potential of self-supervised methods in remote sensing change detection.

III. Methodology

A. Overall Architecture

The overall structure of the proposed method TD-SSCD is shown in Fig. 2. The framework consists of two stages, including stage 1: learning the temporal invariance features from bitemporal images and stage 2: self-supervised learning and optimization based on the temporal and differential. It is assumed that a set of bitemporal patches are extracted from the images and the corresponding DIs. Suppose a set of bitemporal patches are extracted from the images and the corresponding DIs. In the first stage of self-supervised learning, the bitemporal patches are input into the two branches, respectively, to realize that each branch can learn image information of different temporal. In the second stage, in order to consider the advantages of bitemporal and DIs in change detection in the SSCD framework, this study proposes a new positive feature embedding construction strategy for SSCD, that is, the feature representation F_{ef} (i.e., early fusion) of temporal images and the feature representation F_{lf} (i.e., late fusion) of DIs as

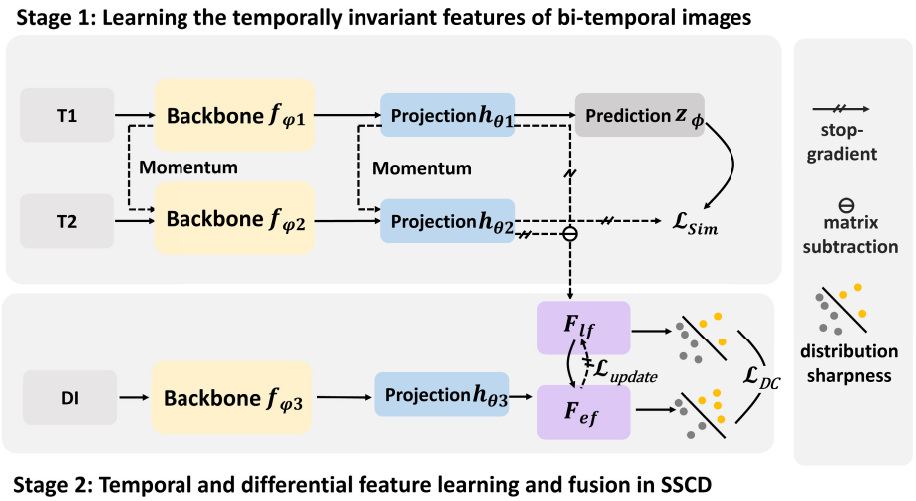

Fig. 2. Schematic overview of the proposed TD-SSCD change detection approach. It can be divided into two main stages. Stage 1: Taking Time1 ($T1$) image as an example, backbone (also referred to as an encoder) $f_{\varphi 1}$ is used to extract semantic features from images, and then, the extracted features are projected into a common domain through an MLP network (i.e., projection head $h_{\theta 1}$). Immediately afterward, the projected features from the projection head $h_{\theta 1}$ are input into the prediction head z_{θ} and converted into a new feature representation. In stage 1, the feature learning is optimized under the two temporal features distance loss $\mathcal{L}_{\text{sim}}$. Stage 2: in the differential branch, the DI feature representations F_{ef} are extracted from the backbone $f_{\varphi 3}$ and the projection heads $h_{\theta 3}$, and the feature representation F_{lf} of the bitemporal image changes in stage 1 is computed. F_{ef} and F_{lf} are utilized to construct a distribution consistency loss $\mathcal{L}_{\text{DC}}$ to train the differential branch. Finally, the two branches can be optimized to mutually enhance each other through an alternating iteration learning strategy (referred to as $\mathcal{L}_{\text{update}}$). Stop-gradient means that gradients cannot be back-propagated through the encoder. Momentum means that a slowly progressing module of $T2$ implemented as a momentum-based moving average of the module of $T1$.

positive feature pairs. Furthermore, to better achieve SSCD learning of these two features, we propose an optimization method with iterative alternation of distribution consistency loss and update loss to guide the bitemporal image branch and DI branch networks to learn from each other. By this mutual learning approach, the information of bitemporal and DIs can be fully exploited, and the loss of important information for determining changes is greatly reduced. In addition, in order to better learn the change information of bitemporal and DIs, we propose a distribution sharpness component to dynamically suppress the error information, highlight the change region, as well as ensure the flexibility of the model.

B. Stage 1: Learning the Temporal Invariance Features of Bitemporal Images

In general, the multiple views in the self-supervised learning methods are obtained by different augmentation techniques, such as cropping, color distortion, or random noise [48], [49]. Fortunately, remote sensing change detection tasks (bitemporal images) naturally have multiple views and do not require data transformations. The difference between the multiple views is mainly due to seasonal factors and imaging conditions. Therefore, this difference can guide the model to learn the temporal invariance features in the first stage.

As seen from stage 1, the two branches have similar structures, i.e., backbone and projection head, where each backbone extracts the useful features from the image at each time, and the projection head is a multilayer perceptron (MLP) network that projects the extracted features into a common domain. In an ideal scenario, the unchanged region would produce similar features from backbones $h_{\theta 1}$ and $h_{\theta 2}$. However, this may cause the output of $h_{\theta 1}$ and $h_{\theta 2}$ to be exactly the same, resulting network collapse. So, we add a two-layer MLP (i.e., prediction head z_{θ}) after $h_{\theta 1}$. These designs have been shown to be effective in BYOL [26], MoCo [41], and SimCLR [42].

Suppose there is a pair of bitemporal images $x_1 \in \mathbb{R}^{H,W,C}$ and $x_2 \in \mathbb{R}^{H,W,C}$, where x_1 and x_2 are the unchanged patches in images $T1$ and $T2$, respectively. The output vectors of the two branches are denoted as $z_1 \triangleq z_{\phi}(h_{\theta 1}(f_{\varphi 1}(x_1)))$ and $p_2 \triangleq h_{\theta 2}(f_{\varphi 2}(x_2))$. Thus, the task of stage 1 can be defined as follows:

$$\theta 1, \varphi 1 = \arg\min_{\theta,\varphi 1}\{\text{sim}(z_1, p_2)\} \tag{1}$$

$$\text{sim}(x, y) = \frac{x}{\|x\|_2} * \frac{y}{\|y\|_2} \tag{2}$$

where sim is a measure of the feature similarity between patch x_1 and x_2. $\|\cdot\|_2$ refers to L_2 normalization.

In addition, we use a strategy without negative pairs to maximize the similarity of the different temporal patches, and at the same time, to prevent the parameters of $h_{\theta 1}$ and $f_{\varphi 1}$ and $h_{\theta 2}$ and $f_{\varphi 2}$ from being exactly the same, we set the parameters of $h_{\theta 2}$ and $f_{\varphi 2}$ to a constant (stop-gradient [25]) and use the updated parameters of $h_{\theta 1}$ and $f_{\varphi 1}$ to slowly update $h_{\theta 2}$ and $f_{\varphi 2}$ (momentum update [41]). This step of the momentum update can be written as follows:

$$\varphi 2, \theta 2 \leftarrow \lambda\varphi 2 + (1-\lambda)\varphi 1, \lambda\theta 2 + (1-\lambda)\theta 1 \tag{3}$$

where $\lambda \in [0, 1)$ is a momentum coefficient, which is usually set to 0.999 as recommended by [40]. Based on these learning strategies in stage 1, the contrastive loss $\mathcal{L}_{\text{sim}}$ can be expressed as follows:

$$\mathcal{L}_{\text{sim}} = -\left(\frac{1}{2}\text{sim}\left(\hat{F}_z, \text{stopgrad}\left(\tilde{F}_{h2}\right)\right) + \frac{1}{2}\text{sim}\left(\tilde{F}_z, \text{stopgrad}\left(\hat{F}_{h2}\right)\right)\right) \tag{4}$$

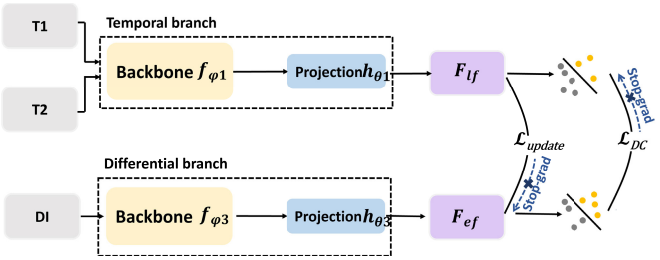

Fig. 3. Details of stage 2 in TD-SSCD.

where $\hat{F}_z$ and $\tilde{F}_z$ represent the final output feature maps of x_1 and x_2 in z_ϕ. $\hat{F}_{h2}$ and $\tilde{F}_{h2}$ represent the final output feature map of x_1 and x_2 in $h_{\theta 2}$. stopgrad indicates the stop-gradient operation.

C. Stage 2: Temporal and Differential Feature Learning and Fusion in SSCD

The pretext process at stage 2 is shown in Fig. 3. The existing SSCD methods generally use features learned from bitemporal images to capture image changes (i.e., late fusion). It aims to extract features from the bitemporal images and detect the changes by measuring the distance between feature pairs. It can reduce pseudochanges, but it can also lead to loss of the original image change information. In contrast, the change detection based on the differential map (i.e., early fusion) focuses more on the information of surface changes [51]. However, the difference map may contain pseudochanges that are caused by seasonal, insolation, and atmospheric features. In summary, both late and early fusions have their advantages and disadvantages. Therefore, in order to combine their advantages in SSCD, we propose a novel self-supervised strategy. In general, we propose new positive embedding strategy, i.e., F_{ef} and F_{lf} (see Figs. 2 and 3). Meanwhile, to achieve deep integration and complementarity of F_{ef} and F_{ef} in the SSCD framework, we propose an alternating iterative training method, including distribution consistency loss and update loss, to guide the mutual learning between temporal and differential branches. In the proposed learning strategy, a branch in the fixed framework is fixed to update another branch, alternating iterative updates, and loops until convergence. In addition, we propose a distribution sharpness component to dynamically suppress misinformation, highlight the regions of change, and ensure model flexibility. The details are as follows.

1) Construction of Positive Embedding Pairs: Conventional self-supervised learning encourages the network to learn the invariance of the transformation by forcing different views of the same image (i.e., positive sample pairs) to be similar. However, this approach is more inclined to obtain the semantic features of the images and cannot mine the temporal change information between bitemporal images effectively. Therefore, we propose a new positive feature embedding pair in the bitemporal differential space for SSCD. In detail, the feature maps (i.e., F_{ef} and F_{lf}) learned by the bitemporal images (late fusion feature) and their DI (early fusion feature) are considered as the positive embedding pairs. In this way, both temporal and differential information can be considered for change detection. To achieve this goal, we add a differential branch to learn the features of the DI and realize the fusion of the differential branch (with $f_{\varphi 3}$ and $h_{\theta 3}$ as the backbone and projection) and temporal branch (with $f_{\varphi 1}$ and $h_{\theta 1}$ as the backbone and projection, respectively) during the network learning (see Fig. 3)

$$F_{\text{ef}} = h_{\theta 3}(f_{\varphi 3}(\text{DI})) \tag{5}$$

$$F_{\text{lf}} = h_{\theta 1}(f_{\varphi 1}(x_1)) - h_{\theta 1}(f_{\varphi 1}(x_2)). \tag{6}$$

2) Alternating Iteration Learning Strategy: In stage 2 of the training process, an alternating learning strategy is proposed to guide the temporal and differential branches to learn from each other. However, at the beginning of training, F_{ef} is unreliable since $f_{\varphi 3}$ and $h_{\theta 3}$ are randomly initialized, so that the well-trained $f_{\varphi 1}$ and $h_{\theta 1}$ may learn the error information from F_{ef}. In turn, F_{lf} generated by $f_{\varphi 1}$ and $h_{\theta 1}$, which has learned the error information, can mislead $f_{\varphi 3}$ and $h_{\theta 3}$. This process can continuously and iteratively transfer the incorrect information between the two branches during the training process and eventually lead to model collapse in both branches. Consequently, this traditional self-supervised training strategy can hardly ensure the interactive learning between the two branches.

To deal with this issue, we propose an alternating iteration learning strategy. To be specific, when completing the training of $h_{\theta 1}$ and $f_{\varphi 1}$, we stop the gradient computation and feedback of $h_{\theta 1}$ and $f_{\varphi 1}$, and the maximum similarity between F_{ef} and F_{lf} is used to guide $h_{\theta 3}$ and $f_{\varphi 3}$ to learn the knowledge from F_{lf}. Afterward, likewise, the differential branch is frozen by fixing $h_{\theta 3}$ and $f_{\varphi 3}$, and the temporal branch (i.e., $h_{\theta 1}$ and $f_{\varphi 1}$) is updated by learning the information from the differential branch (i.e., F_{ef}). In this way, the aforementioned shortcoming caused by the traditional self-supervised learning strategy can be effectively overcome, since the proposed alternating learning strategy can iteratively and progressively strengthen the temporal and differential branches in order and boost their information/knowledge interactions during the training process.

For the proposed alternating learning strategy, we first introduce the training of the differential branch while freezing the temporal branch. It should be noted that self-supervised learning usually calculates the similarity based on the feature distances. However, simply narrowing the feature distance between multiple views cannot express and model the complex correlation between F_{ef} and F_{lf}. In the remote sensing change detection, F_{ef} and F_{lf} should have similar class probability distributions since they both aim to distinguish between changed and unchanged areas. Therefore, in this study, we propose to optimize the consistency of the multiple-view probability distributions to guide $h_{\theta 3}$ and $f_{\varphi 3}$ and $h_{\theta 1}$ and $f_{\varphi 1}$ to learn from each other. The symmetrized distribution consistency (dc) can be written as follows:

$$\text{DC}(F_{\text{lf}}, F_{\text{ef}}) = \frac{1}{2}\text{H}(F_{\text{lf}}|F_{\text{ef}}) + \frac{1}{2}\text{H}(F_{\text{ef}} \mid F_{\text{lf}}) \tag{7}$$

where H(|) refers to the conditional entropy. Particularly, since during the alternating learning, the gradient feedback of $h_{\theta 1}$ and $f_{\varphi 1}$ is stopped, and the overall distribution consistency

loss can be defined as follows:

$$\mathcal{L}_{\mathrm{DC}} = \mathrm{dc}(\mathrm{stopgrad}(F_{\mathrm{lf}}), F_{\mathrm{ef}}). \tag{8}$$

Afterward, the temporal branch is updated while freezing the differential branch. In order to update the frozen $h_{\theta 1}$ and $f_{\varphi 1}$, an update strategy is designed. In the self-supervised contrastive learning, momentum update is usually adopted to gradually transfer the knowledge when updating the frozen network. However, this update strategy is based on the assumptions that the data inputs to both encoders are different views of the same image or that the differences between the two encoders are small. Unfortunately, significant differences exist between the temporal and differential encoders in stage 2. More specific, $h_{\theta 1}$ and $f_{\varphi 1}$ aim to extract the semantic features from the temporal images, while $h_{\theta 3}$ and $f_{\varphi 3}$ aim to learn the change information from the DI. This violates the aforementioned assumptions of the commonly used update strategy. To overcome this issue, we propose a frozen network update strategy based on knowledge distillation, which can be expressed as follows:

$$\mathrm{KD}(x, y) = -\left(\sum_{j}^{N} p(y/\tau)\log(p(x/\tau)\right) \tag{9}$$

where $p(x) = \exp(x_j)/\sum_k^N \exp(x_k)$, and τ is a scalar temperature parameter. This strategy can transfer the knowledge from the differential branch to the temporal branch, making the latter more focus on the change information. Therefore, the update loss $\mathcal{L}_{\mathrm{update}}$ can be defined as follows:

$$\mathcal{L}_{\mathrm{update}} = \frac{1}{2}(\mathrm{KD}(\mathrm{stopgrad}(F_{\mathrm{ef}}), F_{\mathrm{lf}}) + \mathrm{KD}(F_{\mathrm{lf}}, \mathrm{stopgrad}(F_{\mathrm{ef}}))). \tag{10}$$

3) Distribution Sharpness: Although the distribution consistency loss can be an effective guide for $h_{\theta 3}$ and $f_{\varphi 3}$ to understand the feature distribution of F_{lf}, there are still some limitations.

1) In the change detection, due to the low prior probability of the change, in most cases, pixels in the same spatial location and in different time phases are usually unchanged. In this way, the training of the network is more inclined to learn the unchanged sample distribution and ignore the changed samples.
2) Moreover, direct calculation of distribution consistency loss may lead to homogeneity between F_{ef} and F_{lf}.

To overcome these issues, in this study, the distribution sharpness strategy is proposed, as shown in Fig. 4.

First, both F_{ef} and F_{lf} represent the intensity of change, which can be regarded as an initial change map. Therefore, we further divide the pixels of the feature map into two classes (i.e., changed samples ω_{ch} and unchanged samples ω_{un}) according to their change intensity to provide more reliable distribution information. Specifically, we use a dynamic threshold $\mathcal{T}$ to select a small number of samples with a large possibility of change from the same batch as changed samples with high confidence, and the rest of the samples are considered to be unchanged. This threshold $\mathcal{T}$ is defined as follows:

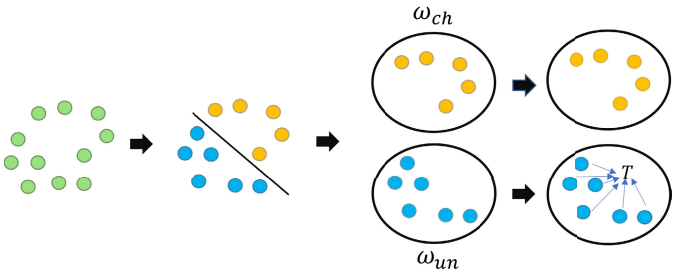

Fig. 4. Illustration of distribution sharpness.

$$\mathcal{T} = \mu + \xi\delta \tag{11}$$

$$w(i, j) \in \begin{cases} \omega_{\mathrm{ch}}, & \text{if } w(i, j) > \mathrm{T} \\ \omega_{\mathrm{un}}, & \text{if } w(i, j) \le T \end{cases} \tag{12}$$

where μ and δ are the mean and the standard deviation of all F_{lf} features in a batch, respectively. Assuming that the proportion of the change regions in a batch is small, μ and δ can measure the F_{lf} feature representation of the unchange/change components, respectively. Thus, ξ is a hyperparameter that balances these two components.

Then, in order to make the model more focused on the change information and encourage the mutual information learning between F_{ef} and F_{lf}, we sharpen the distribution of the samples. For the unchanged sample ω_{un}, we minimize the entropy of the intraclass distribution of each unchanged sample, in order to sharpen the output distribution. For the change sample ω_{ch}, we aim to diversify its distribution to avoid the network assigning all the change samples to the same class. Next, expanding the differences between the changed and unchanged samples. This enables the features of the samples of the unchanged ω_{un} in F_{ef} and F_{lf} more compact and highlights the change ω_{ch}. The processed feature map can be expressed as follows:

$$\tilde{w} = \{\omega_{\mathrm{ch}}, \omega'_{\mathrm{un}}\} \tag{13}$$

$$\hat{w} = \frac{\tilde{w} - \mu_{\tilde{w}}}{\sqrt[2]{\sigma_{\tilde{w}}^2 + \varepsilon}} \tag{14}$$

where $\sigma_{\tilde{w}}^2$ is the standard deviation of $\tilde{w}$, $\hat{w}$ represents the reconstructed feature map, and $\mu_{\tilde{w}}$ is the mean value of $\tilde{w}$.

Finally, the distribution consistency loss of F_{ef} and F_{lf} after distribution sharpness transformation is calculated. This can be defined as follows:

$$\mathcal{L}_{\mathrm{DC}} = \mathrm{DC}(\mathrm{stopgrad}(\hat{w}_{\mathrm{lf}}), \hat{w}_{\mathrm{ef}}) \tag{15}$$

where $\hat{w}_{\mathrm{lf}}$ and $\hat{w}_{\mathrm{ef}}$ denote F_{ef} and F_{lf} after distribution sharpness processing, respectively.

D. Overall Loss and Change Detection

The training process consists of two stages, corresponding to $\mathcal{S}_1$ and $\mathcal{S}_2$ epochs, respectively. For the first $\mathcal{S}_1$ epochs (stage 1), only the contrastive loss $\mathcal{L}_{\mathrm{sim}}$ is used to modulate the network weights. For the subsequent $\mathcal{S}_2$ epochs (stage 2),

Algorithm 1 Algorithm of TD-SSCD Pre-Training

Input:

Bi-temporal images patch: x_1^B, x_2^B, which indicate the image patch in batch B of T1 and T2, respectively.

Differential image patch: DI^B, which indicates the differential image patch in batch B.

Backbone: $f_{\varphi 1}, f_{\varphi 2}, f_{\varphi 3}$, which denote a backbone with parameter $\varphi 1$, $\varphi 2$, $\varphi 3$, respectively.

Projection head: $h_{\theta 1}, h_{\theta 2}, h_{\theta 3}$, which denote a projection head with parameter $\theta 1$, $\theta 2$, $\theta 3$, respectively.

Prediction head: z_ϕ, which denotes a projection head with parameter ϕ

Parameters: $\theta_1, \varphi_1, \theta_2, \varphi_2. \theta_3, \varphi_3, \phi$

Epoch: $\mathcal{S}_1, \mathcal{S}_2$ denote the number of pre-trained epochs at stage 1 and stage 2, respectively, and $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$. The first N epochs at stage 2 is optimized without alternating iteration learning strategy.

Number: n is a counter.

Result: Updated $\theta_1, \varphi_1, \theta_3, \varphi_3$

//Initialization

Randomly initialize parameters θ_1 of $h_{\theta 1}$ and copy to $h_{\theta 2}, h_{\theta 3}$

Randomly initialize parameters φ_1 of $f_{\varphi 1}$ and copy to $f_{\varphi 2}$, $f_{\varphi 3}$

1: **for** $s \leftarrow 1$ to $\mathcal{S}$ **do**
2: **for** $b \in B$ **do**
3: $\hat{F}_z, \tilde{F}_{\tilde{z}} = z_\phi(h_{\theta 1}(f_{\varphi 1}(x_1^b))), z_\phi(h_{\theta 1}(f_{\varphi 1}(x_2^b)))$
4: $\tilde{F}_{h1}, \tilde{F}_{h2} = h_{\theta 2}(f_{\varphi 2}(x_1^b)), h_{\theta 2}(f_{\varphi 2}(x_2^b))$
5: $F_{\text{ef}} = h_{\theta 3}(f_{\varphi 3}(DI^b))$
6: $F_{\text{lf}} = h_{\theta 1}(f_{\varphi 1}(x_1^b)) - h_{\theta 1}(f_{\varphi 1}(x_2^b))$
7: **end for**
8: Calculate contrastive loss $\mathcal{L}_{sim}$
9: Calculate distribution consistency loss $\mathcal{L}_{DC}$
10: Calculate update loss $\mathcal{L}_{update}$
//Stage 1: Learning the Temporal Invariance Features of Bi-Temporal Images
11: **if** $s \leq \mathcal{S}_1$ **then**
12: Update θ_1, φ_1 to minimize $\mathcal{L}_{sim}$
13: Update θ_2, φ_2 with momentum update
//Stage 2: Fusing Temporal and Differential Information
14: **elif** $s \leq \mathcal{S}_1 + N$ **then**
15: Update θ_3, φ_3 to minimize $\mathcal{L}_{DC}$
//Balanced combination of loss functions
16: **elif** $s = \mathcal{S}_1 + N + 2\text{n}$ **then** //n $= 1, 2, 3 \ldots\ldots$
17: Update θ_1, φ_1 to minimize $\mathcal{L}_{update}$
18: **else**
19: Update θ_3, φ_3 to minimize $\mathcal{L}_{DC}$
20: **end if**
21: **end for**
22: **return** parameters $\theta_1, \varphi_1, \theta_3, \varphi_3$

we combine the update loss with the distribution consistency loss to guide F_{ef} and F_{lf} to learn from each other

$$\mathcal{L}_{\text{stage2}} = \begin{cases} \mathcal{L}_{\text{DC}}, & \text{if } \mathcal{S}_2 < N \\ \mathcal{L}_{\text{update}}, & \text{if } (\mathcal{S}_2 < N) = 2n \\ \mathcal{L}_{\text{DC}}, & \text{Otherwise} \end{cases} \tag{16}$$

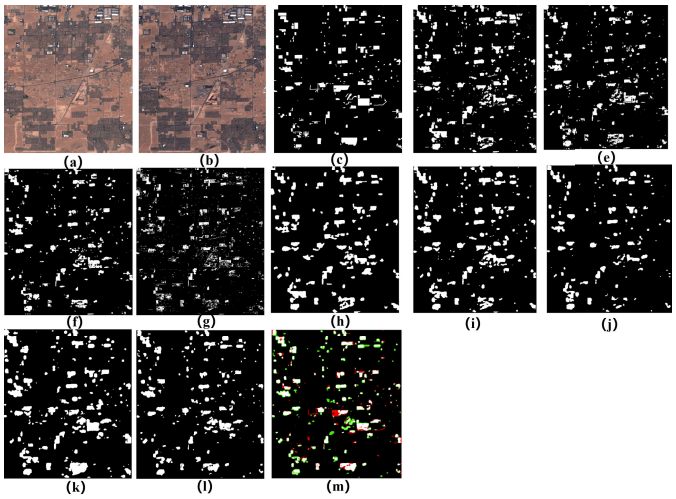

Fig. 5. Visual comparison of the change detection maps obtained by different approaches on the OSCD lasvegas dataset. (a) $T1$ image. (b) $T2$ image. (c) Ground-truth map. (d) FC-EF. (e) SiamUnet conc. (f) SiamUnet diff. (g) PCA-k-means. (h) DCV A. (i) BYOL. (j) SimSiam. (k) SimCLR. (l) Proposed TD-SSCD approach. (m) Confusion map of the proposed TD-SSCD approach (TP: white; TN: black; FP: green; FN: red).

where n is a counter starting at 0. The pseudocode of our proposed TD-SSCD algorithm is shown in the following. Once the network is trained, we can obtain two feature extractors ($h_{\theta 1}$ and $f_{\varphi 1}$ for $T1$ and $T2$, and $h_{\theta 3}$ and $f_{\varphi 3}$ for DI) and generate a feature vector from the image. In detail, given an input image $x \in \mathbb{R}^{H,W,C}$, we can intercept a square local image region with a side length l centered on the pixels in row r and column c and then obtain the feature vector $\mathcal{D}(r, c)$ by a trained feature extractor. We define $\mathcal{D}1(r, c)$, $\mathcal{D}2(r, c)$, and $\mathcal{D}I(r, c)$ as the feature vectors of the bitemporal images and the difference image in row r and column c, respectively. The change intensity map is defined as the difference $e(r, c)$ between the feature vectors

$$e(r, c) = (\|\mathcal{D}1(r, c) - \mathcal{D}2(r, c)\|_2 + \mathcal{D}I(r, c))/2 \tag{17}$$

where the first item represents the change intensity learned from the temporal branch, and the second one stands for the change information derived from the differential branch. Change detection is then realized by setting a suitable threshold value.

IV. Experimental Results and Discussion

A. Dataset Description

We used two datasets to validate the proposed TD-SSCD method (Figs. 5–8): the Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI) dataset [52] and the Onera Satellite Change Detection (OSCD) dataset [53]. The SZTAKI dataset contains 13 image pairs (acquired between 2000 and 2005) provided by The Institute of Geodesy, Cartography and Remote Sensing in Hungary. This dataset consists of three parts, including Archive (one pair), SZADA (seven pairs), and Tiszadob (five pairs). The spatial resolution of the imagery is 1.5 m/pixel. The OSCD dataset (24 image pairs) was created for change detection using Sentinel-2 images acquired between 2015 and 2018. The dataset consists of 24 pairs

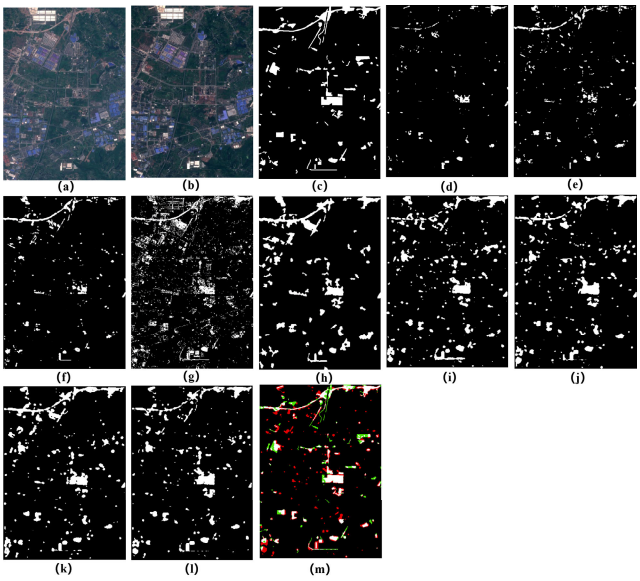

Fig. 6. Visual comparison of the change detection maps obtained by different approaches on the OSCD chongqing dataset. (a) $T1$ image. (b) $T2$ image. (c) Ground-truth map. (d) FC-EF. (e) SiamUnet conc. (f) SiamUnet diff. (g) PCA-k-means. (h) DCV A. (i) BYOL. (j) SimSiam. (k) SimCLR. (l) Proposed TD-SSCD approach. (m) Confusion map of the proposed TD-SSCD approach (TP: white; TN: black; FP: green; FN: red).

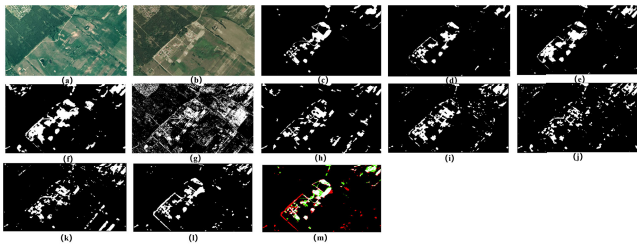

Fig. 7. Visual comparison of the change detection maps obtained by different approaches on the SZADA-2 dataset. (a) $T1$ image. (b) $T2$ image. (c) Ground-truth map. (d) FC-EF. (e) SiamUnet conc. (f) SiamUnet diff. (g) PCA-k-means. (h) DCV A. (i) BYOL. (j) SimSiam. (k) SimCLR. (l) Proposed TD-SSCD approach. (m) Confusion map of the proposed TD-SSCD approach (TP: white; TN: black; FP: green; FN: red).

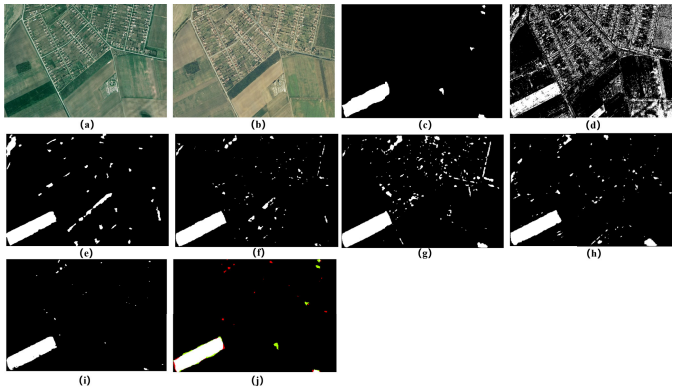

Fig. 8. Visual comparison of the change detection maps obtained by different approaches on the Tiszadob-5 dataset. (a) $T1$ image. (b) $T2$ image. (c) Ground-truth map. (d) PCA-k-means. (e) DCVA. (f) BYOL. (g) SimSiam. (h) SimCLR. (i) Proposed TD-SSCD approach. (j) Confusion map of the proposed TD-SSCD approach (TP: white; TN: black; FP: green; FN: red).

of multispectral images (10 m/pixel). In the experiments, the RGB and near-infrared bands were used. This dataset has two parts, including test set (ten pairs) and train set (14 pairs).

B. Comparison Methods

To verify the effectiveness of the proposed TD-SSCD method, we compared it to a series of SOTA unsupervised, supervised, and SSCD methods.

1) PCA-k-means [31]: an unsupervised change detection method that extracts feature vectors with PCA and distinguishes between changed and unchanged regions through k-means.
2) DCVA [16]: an unsupervised deep learning change detection method that adopts a pretrained CNN to extract the spatial-contextual information and then uses CVA to identity changed pixels.
3) BYOL [28]: it utilizes the self-supervised BYOL method to obtain the feature representation of bitemporal images and uses appropriate thresholds to distinguish changed and unchanged regions.
4) SimSiam [25]: a self-supervised learning method and its basic idea is to force positive samples to be similar. In this study, it was carried out for change detection.
5) SimCLR [42]: a self-supervised learning method that learns a feature representation by forcing the positive samples to be similar and the negative samples to be dissimilar.
6) FC-EF [47]: a supervised deep learning change detection method that concatenates bitemporal images before passing them through the network. The change map can be obtained by conducting the "encoded–decoded" on the fused image.
7) FC-Siam-conc [47]: a supervised deep learning change detection method that uses a Siamese encoder to obtain the features of bitemporal images and concatenates them in the decoding step.
8) FC-Siam-diff [47]: a supervised deep learning change detection method that obtains the features of bitemporal images through a Siamese extractor, and the absolute value of the difference between the bitemporal image features is connected in the decoding step.

C. Experimental Settings

The proposed TD-SSCD method was implemented in PyTorch, with the training and test of the network on an NVIDIA RTX 2080Ti GPU. The patch size of the input data was 8, and patches were extracted from the scenes with a stride of 4. ResNet-18 [54] was adopted as the backbone of the proposed framework. In stage 1, a stochastic gradient descent (SGD) optimizer was applied for 500 epochs, with the learning rate set to 0.03. The encoder with the lowest loss in stage 1 training process was saved for stage 2 task. In stage 2, we used the SGD optimizer for 20 epochs, with the learning rate set to 0.03. N and τ were set to 10 and 0.5, respectively. ξ is set to 3.

For our experiments on SSCD, we selected the train set portion (14 pairs) of the OSCD dataset for pretraining (without using the labels of the datasets) and selected OSCD_lasvegas and OSCD_chongqing [28] of them to test the model performance according to the recommendations of the existing literature. For the SZTAKI dataset, we selected images from

the SZADA and Tiszadob regions (excluding the Tiszadob-5 and SZADA-2, total ten pairs) to pretrain our model (without considering the labels of the datasets) and used Tiszadob-5 and SZADA-2 [54] from these two regions according to the recommendations of the existing literature to test the performance of the model.

In the experiments for the supervised change detection, for the OSCD dataset, according to [53], 14 of the images (training set for OSCD dataset) were used for training, and the OSCD_lasvegas and OSCD_chongqing regions were used for testing. With regard to the SZTAKI dataset, in terms of [47], in the SZADA region, the SZADA-2 image pair was used as test, and the other six pairs of images in this region were used for training. In the Tiszadob region, the Tiszadob-5 image pair was used for testing, and the remaining four image pairs were used for training.

To quantitatively assess the methods, three evaluation metrics are used in this article: overall accuracy (OA), $F1$-score ($F1$), and kappa coefficient (kappa). The higher the values of these metrics, the better the model performance. The three evaluation metrics can be calculated as follows:

$$\begin{aligned} \mathrm{OA} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \\ F1 &= \frac{2\times \mathrm{precision}\times \mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} \\ \mathrm{Kappa} &= \frac{\mathrm{OA}-\mathrm{PE}}{1-\mathrm{PE}} \\ \mathrm{PE} &= \frac{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2} \\ &\quad + \frac{(\mathrm{TN}+\mathrm{FN})\times(\mathrm{FP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})^2} \end{aligned} \tag{18}$$

where FN, TP, FP, and TN refer to false negative, true positive, false positive, and true negative, respectively.

D. Results and Comparison

1) OSCD_lasvegas: Fig. 5 shows the change detection maps obtained by the comparison methods and the proposed TD-SSCD method. It is clear that the change detection results of the traditional method (i.e., PCA-k-means) differ from those of the other ones, and its results exhibit significant salt-and-pepper noise. In contrast, this issue can be better addressed by DCVA. However, DCVA pays more attention to the large changed areas and misses a lot of small discrete changed areas. The supervised and self-supervised methods can effectively suppress the false detections and accurately detect the small changed areas.

Table I lists the experimental results obtained on the OSCD_lasvegas dataset, with the best performance highlighted in bold. It can be seen that most of the change detection methods can achieve a good accuracy. As expected, the supervised methods (i.e., FC-EF, SiamUnet_conc, and SiamUnet_diff) show a better performance than the unsupervised ones in most of the metrics. It is interesting to see that the SSCD methods are superior to some of the supervised methods (e.g., FC-EF). In particular, the proposed TD-SSCD model shows the best performance among the unsupervised methods, with the $F1$ (72.11%), OA (95.38%), and kappa (69.60%) all being the highest. For the other self-supervised methods, they show similar performances in most of the evaluation metrics. In addition, we can find that the self-supervised methods bring great improvements in kappa and $F1$, compared with DCVA. This shows that the feature extractor obtained by the self-supervised learning is superior to the pretrained feature extractor used by DCVA. It should be kept in mind that the supervised learning approach uses a large number of labeled samples for training, while the self-supervised learning methods train the networks based on the explicit information of the images but without the labels. In our experiments, the self-supervised methods reach or even surpass the accuracy of the SOTA supervised learning methods, which illustrates the great potential of self-supervised learning in change detection tasks.

TABLE I
QUANTITATIVE EVALUATION OF THE DIFFERENT APPROACHES APPLIED TO THE OSCD_LASVEGAS DATASET

Type	Method	F1	OA	Kappa
Supervised	FC-EF	0.6466	0.9355	0.6121
	SiamUnet_conc	0.7093	0.9553	0.6852
	SiamUnet_diff	**0.7292**	**0.9583**	**0.7066**
Unsupervised	PCA-k-means	0.6212	**0.9454**	0.5919
	DCVA	**0.6321**	0.9318	**0.5957**
Self-supervised	BYOL	0.6627	0.9387	0.6299
	SimSiam	0.6507	0.9488	0.6232
	SimCLR	0.6732	0.9344	0.6390
	Proposed	**0.7211**	**0.9538**	**0.6960**

TABLE II
QUANTITATIVE EVALUATION OF THE DIFFERENT APPROACHES APPLIED TO THE OSCD_CHONGQING DATASET

Type	Method	F1	OA	Kappa
Supervised	FC-EF	0.4165	0.9474	0.4416
	SiamUnet_conc	**0.4988**	**0.9520**	**0.4756**
	SiamUnet_diff	0.4917	0.9488	0.4660
Unsupervised	PCA-k-means	0.4723	0.9161	0.4295
	DCVA	**0.4896**	**0.9206**	**0.4488**
Self-supervised	BYOL	0.5588	0.9321	0.5238
	SimSiam	0.5577	0.9360	0.5241
	SimCLR	0.5612	0.9348	0.5272
	Proposed	**0.5941**	**0.9383**	**0.5622**

TABLE III
QUANTITATIVE EVALUATION OF THE DIFFERENT APPROACHES APPLIED TO THE SZADA-2 DATASET

Type	Method	F1	OA	Kappa
Supervised	FC-EF	**0.6902**	**0.9674**	**0.6731**
	SiamUnet_conc	0.6881	0.9583	0.6662
	SiamUnet_diff	0.5967	0.9291	0.5628
Unsupervised	PCA-k-means	0.2932	0.8111	0.2228
	DCVA	**0.4662**	**0.9261**	**0.4276**
Self-supervised	BYOL	0.5788	0.9395	0.5476
	SimSiam	0.4430	0.9318	0.4430
	SimCLR	0.4945	0.9302	0.4585
	Proposed	**0.6516**	**0.9571**	**0.6288**

2) OSCD_chongqing: The results of the comparison between the proposed TD-SSCD model and the other comparison methods on this region are shown in Fig. 6 and Table II. As can be seen from Fig. 6, there are many missed detections in the FC-EF results, which indicates that the information contained in the difference image alone is not sufficient for change detection. In contrast, the proposed TD-SSCD method does not present this kind of omission errors, although it also uses a difference image. Meanwhile, compared with the other self-supervised methods, the proposed TD-SSCD method successfully eliminates some pseudo-changes and obtains more accurate and clearer results. This demonstrates that the proposed TD-SSCD method reasonably combines the advantages of original bi-temporal images and their difference

information during the change feature learning. As can be seen in Table II, the proposed TD-SSCD model obtains the best performance among all the methods (even compared with the supervised ones) in terms of $F1$ (59.41%), OA (93.83%), and kappa (56.22%). Compared with the other methods, the proposed model obtains improvements of 12.1% (over FC-EF), 8.7% (over SiamUnet_conc), 9.6% (over SiamUnet_diff), 13.3% (over PCA-k-means), 11.3% (over DCVA), 3.8% (over BYOL), 3.8% (over SimSiam), and 3.5% (over SimCLR), respectively. Note that the accuracy of SimCLR is next to the proposed TD-SSCD, and a possible reason is that it uses a positive and negative pair strategy, which makes it more sensitive to small changes.

3) SZADA-2: The SZADA dataset contains many kinds of changes, including roadway construction, building construction, new cultivated land, and so on. The change detection results of the different methods on the SZADA-2 dataset are shown in Fig. 7. It can be seen that the results of all the unsupervised methods have both omissions and false alarms to some degree. The proposed TD-SSCD method obtains a change map that is more consistent with the ground truth. As shown in Table III, the proposed TD-SSCD model obtains the best overall performance among the unsupervised and self-supervised schemes. In particular, compared with the BYOL model, the improvements of kappa, OA, and $F1$ achieved by the proposed TD-SSCD method are 8.4%, 1.42%, and 6.13%, respectively. It should be noted that the performance of different self-supervised networks is significantly different in this experiment, unlike the OSCD experiments. This suggests that the study areas have a significant impact on the results of the self-supervised methods. However, it is promising that the proposed TD-SSCD method consistently shows a superior performance in both datasets [i.e., OSCD (10 m) and SZTAKI (1.5 m)]. For the three supervised learning methods, it is clear that the performance of SiamUnet_diff is worse than that of the other two. FC-EF obtains the highest accuracy among the supervised methods. This reflects the importance of difference images in change detection since it can properly guide the network to learn the change information from the bitemporal images. This also illustrates the rationality of the proposed strategy, i.e., fusion of temporal and DIs for self-supervised learning of change detection.

4) Tiszadob-5: Due to the small number of training samples in this dataset, it is difficult to obtain a desirable performance with the supervised methods. This illustrates the necessity of the unsupervised change detection methods. The visual results of different methods on the Tiszadob-5 dataset are shown in Fig. 8. As can be seen, the main land-cover types in the region are farmland and grassland, with the changes mainly appearing in the vegetation. This leads to a large number of false alarms in the comparison methods. For instance, the traditional method (i.e., PCA-k-means) is heavily influenced by noise, and a large amount of unchanged vegetation is mistakenly identified as changed area. However, the proposed TD-SSCD method successfully eliminates these false changes and produces more accurate and clearer results. Table IV lists the quantitative evaluation results, which are consistent with the visual inspection. The proposed TD-SSCD model obtains the best accuracy, and compared with the other self-supervised methods, TD-SSCD shows clear superiority. This indicates that the fusion of temporal and DIs enhances the ability of the self-supervised learning of the change information. In this way, the robustness of the model to both omission and commission errors is enhanced.

TABLE IV
QUANTITATIVE EVALUATION OF THE DIFFERENT APPROACHES APPLIED TO THE TISZADOB-5 DATASET

Type	Method	F1	OA	Kappa
Supervised	FC-EF	-	-	-
	SiamUnet_conc	-	-	-
	SiamUnet_diff	-	-	-
Unsupervised	PCA-k-means	0.2375	0.7783	0.1793
	DCVA	**0.6725**	**0.9666**	**0.6556**
Self-supervised	BYOL	0.7916	0.9805	0.7815
	SimSiam	0.7176	0.9692	0.7024
	SimCLR	0.7648	0.9771	0.7531
	Proposed	**0.8951**	**0.9915**	**0.8906**

E. Ablation Studies

We further investigate the role of each component in the TD-SSCD (see M1–M8). In this section, M1–M6 contain information from the differential feature learning and the temporal feature learning branches, and M7 and M8 contain only temporal semantic information.

1) M1: Complete two-stage change detection framework TD-SSCD. Please kindly note that in stage 2, the loss function is a combination of the distribution consistent loss $\mathcal{L}_{\text{DC}}$ and the two-branch update loss $\mathcal{L}_{\text{update}}$, with an alternating iteration learning strategy, and a distribution sharpness operation that magnifies the difference between changed and unchanged features.
2) M2: Elimination of the update loss $\mathcal{L}_{\text{update}}$ in stage 2 on the basis of M1.
3) M3: Elimination of distribution sharpness operation in stage 2 on the basis of M1.
4) M4: Replacement of the distribution consistent loss $\mathcal{L}_{\text{DC}}$ with the mean square error loss (i.e., a Euclidean distance between two features) on the basis of M3.
5) M5: Elimination of $\mathcal{L}_{\text{update}}$ on the basis of M3.
6) M6: Replacement of the alternating iteration learning strategy in stage 2 with a momentum update strategy compared to M1.
7) M7: Compared to M1–M6, stage 2 was completely removed. That is, information from the differential feature learning branch is completely lost.
8) M8: Elimination of the prediction head on the basis of M7.

Taking OSCD_chongqing dataset as an example, the experimental results are shown in Table V. It can be seen that the performance of the model deteriorates gradually with the removal of the proposed modules. In general, the accuracy of the two-stage models, with the exception of M6, is significantly higher than the performances of the models using only temporal semantic information. It suggested that the DIs provide useful information to correctly identify changing and unchanged regions. Furthermore, compared with M1, the severe drop in accuracy of M6 mainly attributes to the large heterogeneity between the encoders of the differential branch and the temporal branch. In this case, the momentum update strategy is not applicable, and the alternating iteration updating strategy works well.

TABLE V
QUANTITATIVE EVALUATION OF THE ABLATION STUDY RESULTS WITH OSCD_CHONGQING DATASET

Method	Kappa	OA	F1
M1	**0.5622**	**0.9383**	**0.5941**
M2	0.5540	0.9374	0.5864
M3	0.5476	0.9385	0.5797
M4	0.5331	0.9351	0.5668
M5	0.5354	0.9374	0.5682
M6	0.3394	0.9313	0.3753
M7	0.5238	0.9321	0.5588
M8	0.3001	0.8859	0.3560

Meanwhile, in the ablation studies of the two-stage model.

1) Comparative results between M1 and M2, as well as between M3 and M5, it can be seen that the addition of $\mathcal{L}_{\text{update}}$ achieves a positive interaction between the temporal and the differential branches.
2) M3 and M4 verify that $\mathcal{L}_{\text{DC}}$ can more robustly measure the consistency of the two branches in characterizing feature changes than using Euclidean distance-based feature matching.
3) The comparison between M1 and M3 shows that distribution sharpness can be effective in separating changed and unchanged features in differential features.

In the single-stage ablation studies: M8 can hardly correct the change region (kappa: 0.3560), suggesting that the prediction head can effectively prevent the collapse of the temporal branch. Since the learning process of the differential branch (i.e., stage 2) depends on the characterization of the temporal features and the alternating iteration updating in the initial stage, it is considered the prediction head to be indispensable in the temporal branch.

F. Effect of the Amount of Self-Supervised Data

Since the number of pretraining samples is one of the keys to the performance of a self-supervised method, this section analyzes the relationship between the performance of different SSCD methods and the number of pretraining samples. We conducted experiments on OSCD dataset and designed two pretraining dataset scenarios. Specifically, 1) pretraining using the train set of OSCD dataset (ten pairs) and 2) pretraining the train set of OSCD with S2MTCP dataset [56] (1535 pairs). It is worth mentioning that S2MTCP dataset contains 1521 bitemporal Sentinel-2 image pairs for urban areas, with a size of 600 × 600. This dataset is used to augment the amount of data in OSCD dataset for the self-supervised methods in the pretraining phase.

From Table VI, it can be found that with the increase of pretraining data, the performance of most self-supervised methods tends to increase on the whole. Therefore, using a larger dataset for pretraining is beneficial for self-supervised methods. Furthermore, it can be seen that the proposed self-supervised method has obtained good results in the 14-pair image condition, which indicates that even with not very large dataset, the proposed self-supervised learning can make full use of the unlabeled data to obtain good performance. It is worth mentioning that our proposed method consistently reaches the highest performance regardless of the amount of pretraining data, which demonstrates the superiority of TD-SSCD.

G. Influence of Noise

During the imaging process, the signal or image usually has various degradation, noise effects, or variability [57]. Therefore, referring to the setup in [57], we performed randomly degradation of the OSCD dataset to explore the impact. Specifically, we randomly selected 30% of the pixels in the pretraining dataset and test area data of OSCD and added 25 dB of white Gaussian noise into each band of these pixels. The results are shown in Table VII.

TABLE VI
RESULTS OF SELF-SUPERVISED METHODS ON DIFFERENT PRETRAINING DATASETS

Pre-train image number	Model	OSCD_Chognqing		OSCD_Lasvegas	
		F1	Kappa	F1	Kappa
14(the train set of OSCD)	BYOL	0.5588	0.5238	0.6627	0.6299
	SimSiam	0.5577	0.5241	0.6507	0.6232
	SimCLR	0.5612	0.5272	0.6732	0.6390
	TD-SSCD	**0.5941**	**0.5622**	**0.7211**	**0.6960**
1535(S2MTCP +the train set of OSCD)	BYOL	0.5651	0.5312	0.6605	0.6367
	SimSiam	0.5644	0.5316	0.6590	0.6261
	SimCLR	0.5705	0.5381	0.6651	0.6415
	TD-SSCD	0.5969	0.5652	0.7239	0.6992

TABLE VII
SENSITIVITY ANALYSIS OF TD-SSCD WITH DATA DEGRADATION

Model	OSCD_Chongqing		OSCD_Lasvegas	
	F1	Kappa	F1	Kappa
TD-SSCD	0.5941	0.5622	0.7211	0.6960
TD-SSCD (Add Noise)	0.5897	0.5614	0.7256	0.7018

As can be seen from Table VII, the influence of white Gaussian noise on TD-SSCD is not significant. It indicates that our model is still effective in learning the true sematic features after the data degradation.

H. Influence of Land Cover Change in Pretraining Data

In self-supervised learning, multiple views are typically obtained using various augmentation techniques, such as cropping, color distortion, or random noise. However, in remote sensing change detection tasks, it is common to use bitemporal images as multiple views without applying data transformations. In this scenario, the model is guided to learn temporal invariance features by understanding the temporal differences between these multiple views caused by factors, such as seasons and imaging conditions. Unfortunately, the bitemporal images may not only contain temporal differences but also change in land cover types, which can potentially limit the performance of model. To investigate the impact of land cover type changes on self-supervised training, we modified the OSCD dataset by masking the changed regions within it, denoted as TD-SSCD (mask change). The experimental results are presented in Table VIII. From the results, we can find that the performance of the model is improved overall after masking the changed regions. Notably, when comparing the results of the OSCD_chongqing, the improvement in model performance was particularly significant. This suggests that this approach may be more beneficial when performing self-supervised training on a dataset where no land cover type changes have occurred.

I. Influence of ξ

An important hyperparameter of the distribution sharpness module is ξ. When ξ is larger, the change features are more significant and vice versa. To investigate the influence of ξ, we set it to 1, 2, 3, and 4 and observe the change in TD-SSCD performance on OSCD_chongqing

TABLE VIII
SENSITIVITY ANALYSIS OF TD-SSCD IN LAND COVER CHANGE

Model	OSCD_Chongqing		OSCD_Lasvegas	
	F1	Kappa	F1	Kappa
TD-SSCD	0.5941	0.5622	0.7211	0.6960
TD-SSCD (Mask Change)	0.5988	0.5705	0.7231	0.6983

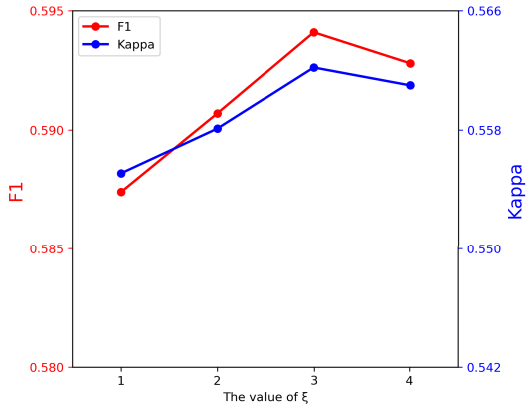

Fig. 9. Classification accuracy versus ξ on OSCD_chongqing.

dataset (see Fig. 9). It can be seen that there is good detection performance on OSCD_chongqing when the value ξ is within a certain range. It benefits from the distribution sharpness module, which makes the model focus on the true change information and filter out nonrelevant information. While when ξ is too large, the distribution sharpness module retains only a small amount of change features information with high confidence, resulting in insufficient information retained to identify the changed and unchanged regions, and the performance of the model decreases. In this experiment, TD-SSCD reaches best result when $\xi = 3$. Therefore, in this article, without loss of generality, we set ξ to 3 in all experiments.

J. Influence of Threshold Method

It is one of the key steps in unsupervised change detection to analyze the change map and determine the change threshold to distinguish the changed and unchanged pixels. To give a fair comparison with current self-supervised learning-based change detection method [28], in this study, we used the adaptive threshold method suggested by [28] to determine appropriate threshold.

To further investigate the influence of threshold method on the performance of TD-SSCD, we also used the Otsu [58] method, which is popular among unsupervised change detection methods, for comparison experiments. It can be found in Tables IX and X that Otsu threshold method gives the model the best accuracy in the OSCD_chongqing dataset, while adaptive threshold method has the highest accuracy in the other datasets. In addition, it can be found that the results of the two threshold methods show small differences, which indicates that our proposed TD-SSCD method is robust to different threshold methods.

K. Analysis of Efficiency

Table XI shows the number of parameters and the computation costs of TD-SSCD, BYOL, SimSiam, and SimCLR on OSCD dataset. It is noted that the table shows the computation cost of these self-supervised methods during the pretraining process. All model runtimes are based on an 11-GB RTX 2080ti GPU. Obviously, SimSiam is the fastest because it uses Siamese network to reduce the parameters and only calculates the similarity between positive pairs to learn the representations. Although SimCLR is also a Siamese network, it is limited by the need to calculate the distance between negative pairs and positive pairs, resulting in the greatest time consumption. The TD-SSCD has a large number of parameters, mainly because it uses multiple feature extractors to learn temporal and differential information, respectively. However, due to its unique alternating iteration learning strategy, the speed is similar to the time spent on BYOL.

TABLE IX
VARIATION OF RESULT AS THRESHOLD METHOD IS VARIED ON OSCD DATASET

Threshold	OSCD_chongqing		OSCD_lasvegas	
	F1	Kappa	F1	Kappa
Adaptive	0.5941	0.5622	0.7211	0.6960
Otsu	0.5883	0.5589	0.7069	0.6812

TABLE X
VARIATION OF RESULT AS THRESHOLD METHOD IS VARIED ON SZTAKI DATASETS

Threshold	SZADA-2		Tiszadob-5	
	F1	Kappa	F1	Kappa
Adaptive	0.6516	0.6288	0.8951	0.8906
Otsu	0.6376	0.6113	0.8898	0.8850

TABLE XI
COMPARISON OF PARAMETERS AND COMPUTATIONAL COSTS OF DIFFERENT METHODS ON THE OSCD DATASET

	GFLOPs	**Pra (k.)**	**Time(min.)**
BYOL	1.506	367.552	51.69
SimSiam	0.757	185.856	24.38
SimCLR	0.757	185.856	91.17
TD-SSCD	2.247	545.088	53.42

V. CONCLUSION

In this article, we have proposed a two-stage self-supervised network for remote sensing change detection, namely, TD-SSCD, by simultaneously considering the temporal images and their differential features in the self-supervised learning. Unlike traditional self-supervised methods, a new stage was proposed during the pretraining process, in order to fully exploit the semantic and change information of unlabeled bitemporal images. Specifically, we designed a balanced loss combination in the new stage to help the two feature representation schemes, i.e., features learned from bitemporal images and their DI, to learn from and boost each other in an alternating manner through the two proposed loss functions (i.e., distribution consistency loss and update loss). Furthermore, a feature distribution sharpness component was designed in terms of the distribution of the two features to improve the learning ability of the model.

The experimental results show that compared with the current SOTA unsupervised and SSCD methods, the proposed TD-SSCD change detection method performed better on two popular change detection datasets and showed a good learning ability for changed areas. In addition, the results showed that the proposed method had similar and even better performance compared to supervised learning change detection methods, which narrowed the gap between unsupervised and supervised change detection.

Frankly, the TD-SSCD method proposed in this article has the following limitations. 1) It requires a certain amount of

unlabeled images to ensure a sufficient model training. 2) We use temporal differences between bitemporal remote sensing images to guide the model to learn temporal invariance features. These temporal differences are mainly caused by factors, such as seasons and imaging conditions. However, in change detection tasks, the bitemporal images contain more than just this temporal difference and the land cover type may change, which may mislead the model learning. Therefore, our future work will focus on building reliable remote sensing image pairs to advance the development of SSCD methods in remote sensing. In addition, we will further study semantic change detection, which can reveal semantic classes before and after the change for more detailed analysis.

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] J. Yang and X. Huang, "30 m annual land cover and its dynamics in China from 1990 to 2019," *Earth Syst. Sci. Data*, vol. 2021, pp. 1–29, Apr. 2021.

[2] 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.

[3] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy, "Land-cover change detection using multi-temporal MODIS NDVI data," *Remote Sens. Environ.*, vol. 105, no. 2, pp. 142–154, Nov. 2006.

[4] C. A. Mucher, K. T. Steinnocher, F. P. Kressler, and C. Heunks, "Land cover characterization and change detection for environmental monitoring of pan-Europe," *Int. J. Remote Sens.*, vol. 21, nos. 6–7, pp. 1159–1181, Jan. 2000.

[5] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2403–2420, May 2010.

[6] X. Huang, Y. Cao, and J. Li, "An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images," *Remote Sens. Environ.*, vol. 244, Jul. 2020, Art. no. 111802.

[7] D. Wen et al., "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 68–101, Dec. 2021.

[8] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

[9] J. Li, X. Huang, and X. Chang, "A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 1–17, May 2020.

[10] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, Jun. 2013.

[11] A. M. El Amin, Q. Liu, and Y. Wang, "Convolutional neural network features based change detection in satellite images," *Proc. SPIE*, vol. 10011, Jul. 2016, Art. no. 100110W.

[12] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, p. 1382, Jun. 2019.

[13] J. Li, B. Zhang, and X. Huang, "A hierarchical category structure based convolutional recurrent neural network (HCS-ConvRNN) for land-cover classification using dense MODIS time-series data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, Apr. 2022, Art. no. 102744.

[14] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[15] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[16] 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.

[17] Y. Li, L. Zhou, G. Lu, B. Hou, and L. Jiao, "Change detection in synthetic aperture radar images based on log-mean operator and stacked auto-encoder," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3090–3096.

[18] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 45–49, Jan. 2019.

[19] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, p. 1688, May 2020.

[20] M. Gong, X. Niu, T. Zhan, and M. Zhang, "A coupling translation network for change detection in heterogeneous images," *Int. J. Remote Sens.*, vol. 40, no. 9, pp. 3647–3672, May 2019.

[21] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[22] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[23] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[24] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.

[25] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[26] 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.

[27] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4405710.

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

[29] R. G. Negri et al., "Spectral–spatial-aware unsupervised change detection with stochastic distances and support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 2863–2876, Apr. 2021.

[30] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations of satellite image time series," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*. Würzburg, Germany: Springer, Sep. 2020.

[31] 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.

[32] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.

[33] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 50, pp. 131–140, Aug. 2016.

[34] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2009.

[35] X. Huang, L. Gao, R. S. Crosbie, N. Zhang, G. Fu, and R. Doble, "Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning," *Water*, vol. 11, no. 9, p. 1879, Sep. 2019.

[36] X. Huang, X. Han, S. Ma, T. Lin, and J. Gong, "Monitoring ecosystem service change in the city of Shenzhen by the use of high-resolution remotely sensed imagery and deep learning," *Land Degradation Develop.*, vol. 30, no. 12, pp. 1490–1501, Jul. 2019.

[37] 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.

[38] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on PCANet," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1792–1796, Dec. 2016.

[39] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[40] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[41] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.
[42] 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.
[43] 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.
[44] H. Li et al., "Remote sensing images semantic segmentation with general remote sensing vision model via a self-supervised contrastive learning method," 2021, *arXiv:2106.10605*.
[45] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[46] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[47] R. Caye 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.
[48] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–6.
[49] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1182–1191.
[50] 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.
[51] X. Hou, Y. Bai, Y. Li, C. Shang, and Q. Shen, "High-resolution triplet network with dynamic multiscale feature for change detection on satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 103–115, Jul. 2021.
[52] C. Benedek and T. Sziranyi, "A mixed Markov model for change detection in aerial photos with large time differences," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.
[53] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2115–2118.
[54] 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.
[55] X. Tang et al., "An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609715.
[56] M. Leenstra, D. Marcos, F. Bovolo, and D. Tuia, "Self-supervised pre-training enhances change detection in Sentinel-2 imagery," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2021, pp. 578–590.
[57] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.
[58] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

Yang Qu received the B.S. degree in Earth information science and technology from Henan Polytechnic University, 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, and with the Hubei Luojia Laboratory, Wuhan University, Wuhan. 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, computation vision and pattern recognition, and remote sensing images.

Dr. Li is the young Editorial Board Member of Geospatial-Information Science (GSIS) and the Guest Editor of the *Remote Sensing* and *Sustainability* (an open access journal from MDPI). She is a Reviewer for more than 30 international journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, *Remote Sensing of Environment*, and *International Society for Photogrammetry and Remote Sensing*.

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

He is working with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan. He is currently a Full Professor of Wuhan University, Wuhan, where he teaches remote sensing, image interpretation, and so on. He is the Head of the Institute of Remote Sensing Information Processing (IRSIP), School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. He has published more than 200 peer-reviewed articles (SCI papers) in the international journals. His research interests include remote sensing image processing methods and applications.

Prof. Huang has been supported by the New Century Excellent Talents in University from the Ministry of Education of China (2011), the China National Science Fund for Excellent Young Scholars (2015), and The National Program for Support of Top-Notch Young Professionals (2017). He was the recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing (ASPRS), in 2010, the National Excellent Doctoral Dissertation Award of China, in 2012, and the recipient for the John I. Davidson President's Award from ASPRS, in 2018. He was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as the Best Reviewer of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, in 2011. He was the winner of the IEEE GRSS Data Fusion Contest, in 2014 and 2021, respectively. He was the Lead Guest Editor of the Special Issue for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the *Journal of Applied Remote Sensing, Photogrammetric Engineering and Remote Sensing*, and *Remote Sensing*. He was an Associate Editor of the *Photogrammetric Engineering and Remote Sensing* (2016–2019), an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (2014–2020), an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2018–2022), and now serves as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (since 2022). He is also an Editorial Board Member for the *Remote Sensing of Environment* (since 2019).

Dawei Wen received the B.S. degree in surveying and mapping from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, in 2018.

She is currently a Lecturer with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan. Her research interests include urban remote sensing, high-resolution image processing, and change detection.