SSLChange: A Self-Supervised Change Detection Framework Based on Domain Adaptation

Yitao Zhao, *Graduate Student Member, IEEE*, Turgay Celik, *Senior Member, IEEE*, Nanqing Liu, *Member, IEEE*, Feng Gao, *Member, IEEE*, and Heng-Chao Li, *Senior Member, IEEE*

***Abstract*— In conventional remote sensing change detection (RSCD) procedures, extensive manual labeling for bi-temporal images is first required to maintain the performance of subsequent fully supervised training. However, pixel-level labeling for change detection (CD) tasks is very complex and time-consuming. In this article, we explore a novel self-supervised contrastive framework applicable to the RSCD task, which promotes the model to accurately capture spatial, structural, and semantic information through the domain adapter (DA) and the hierarchical contrastive head. The proposed SSLChange framework accomplishes self-learning only by taking a single-temporal sample and can be flexibly transferred to mainstream CD baselines. With self-supervised contrastive learning, feature representation pretraining can be performed directly based on the original data even without labeling. After a certain number of labels are subsequently obtained, the pretrained features will be aligned with the labels for fully supervised fine-tuning. Without introducing any additional data or labels, the performance of downstream baselines will experience a significant enhancement. Experimental results on two entire datasets and six diluted datasets show that our proposed SSLChange improves the performance and stability of CD baseline in data-limited situations. The code of SSLChange is available at https://github.com/MarsZhaoYT/SSLChange**

***Index Terms*— Change detection (CD), domain adaption, hierarchical features, image contrastive learning, remote sensing (RS) images, self-supervised learning (SSL).**

I. Introduction

CHANGE detection (CD) task plays a crucial role in land use and land cover (LULC), which aims at detecting and highlighting the changed regions in bi-temporal and multitemporal remote sensing (RS) image sequences [1].

Received 23 September 2023; revised 27 February 2024 and 12 June 2024; accepted 23 October 2024. Date of publication 1 November 2024; date of current version 14 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62271418 and Grant 61871335 and in part by the Natural Science Foundation of Sichuan Province under Grant 2023NSFSC0030. *(Corresponding author: Heng-Chao Li.)*

Yitao Zhao, Nanqing Liu, and Heng-Chao Li are with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China (e-mail: ytzhao@my.swjtu.edu.cn; lansing163@163.com; lihengchao78@163.com).

Turgay Celik is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China, also with the School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa, and also with the Faculty of Engineering and Science, University of Agder, 4604 Kristiansand, Norway (e-mail: celikturgay@gmail.com).

Feng Gao is with the School of Information Science and Engineering, Ocean University of China, Qingdao 266100, China.

Digital Object Identifier 10.1109/TGRS.2024.3489615

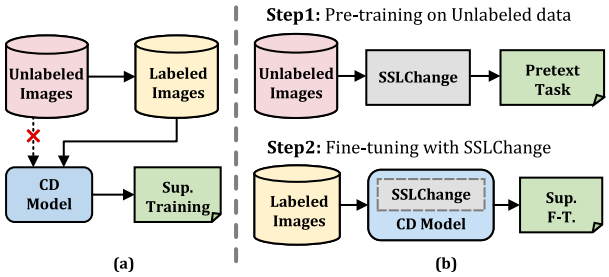

Fig. 1. (a) Previous CD tasks versus (b) our SSLChange. Illustration of the fine-tuning method with a clipped pretrained encoder. F.-T. represents downstream fine-tuning.

Benefiting from the rapid growth of aerospace imaging platforms, a variety of RS images with different resolutions and modalities have greatly facilitated breakthroughs in CD technology. Generally, the CD framework takes registered RS images of the same location acquired at different times as input and identifies changed objects [2]. Considering the wide coverage area of RS images, CD technology automatically and efficiently retrieves changed pixels from large-scale images, greatly reducing the pressure of manual interpretation [3]. Therefore, CD technology is widely applied in urban planning, disaster assessment, environmental monitoring, and other fields [4].

The combination of CD and computer vision has derived numerous excellent algorithms in RSCD tasks [5], [6], [7], [8], [9], [10], [11]. The algorithmic accuracy and efficiency of CD technology have been significantly improved with the dual support of computing power and data resources, promoting a wide range of applications in the fields of land-use planning and street view detection. However, the remarkable successes of such data-driven methods cannot conceal the fatal drawback of severe reliance on large amounts of manual annotations [1]. From the perspective of algorithm strategy, the above-mentioned CD methods belong to the applications of fully supervised learning, in which the deep features extracted from the deep models are compared with the manually annotated ground truth (GT), and then the model is guided to optimize by well-designed loss evaluation metrics. Due to the diversity of RS data format and the requirement for prior knowledge of interpretation, only a few datasets with annotated information are accessible for CD tasks [12]. Data annotation remains a challenging procedure in the RS community. Some researchers proposed several unsupervised CD

algorithms to handle the label-limited constraint and obtained promising results [13], [14], [15], [16], [17]. To alleviate the limitation of insufficient annotated data, a feasible solution is to perform data augmentation on the labeled datasets to enlarge the data volume [18]. Commonly employed data augmentation methods include random cropping and shifting, rotation and flipping, random scaling, brightness adjustment, and noise addition [19], [20]. Another formulation of data augmentation is to perform random masking or mixing on the input patches, such as Mixup [21], Cutout [22], and CutMix [23]. These augmentation methods have sparked remarkable performance in computer vision downstream tasks, but the lack of labeled data remains an obstacle in the RS community.

While fully supervised learning methods have achieved excellent performance, the research interest has gradually shifted to self-supervised learning (SSL) methods. In SSL-based tasks, the model is encouraged to autonomously explore the latent feature from the unlabeled data by constructing pretext tasks. Then, the visual representation ability is transferred to downstream tasks. Representative SSL-based methods proposed for vision tasks like [24], [25], [26], and [27] enable models to obtain satisfying performance on unlabeled datasets, even surpassing fully supervised methods [27]. Researchers have attempted to introduce SSL-based methods in RS vision tasks [28], [29], [30], [31].

However, most of the existing SSL-based RS tasks focus on object-level tasks, and there are only a few explorations on pixel-level processing such as CD tasks [32], [33], [34], [35], [36], [37], [38], [39], [40]. In the absence of pixel-level annotations, it is difficult for existing SSL-based vision models to be immune to various image representation differences in bi-temporal RS images only relying on global feature vector comparison. Specifically, the limitations of existing SSL-based methods on RSCD tasks are as follows.

1) Adaptability of SSL-based frameworks. The mainstream deep learning-based models are not specially designed for pixel-level CD tasks on RS images, ignoring the local spatial features while focusing on global features. In addition, few targeted approaches can adaptively handle the nonlinear gray-scale difference caused by the different climate conditions at bi-temporal imaging moments.
2) Postprocessing procedures. The prediction results of most existing SSL-based CD methods suffer from image noise and uncertain regions. Subsequent postprocessing methods like downstream fine-tuning or threshold segmentation are always utilized to refine the model performance [41], [42]. However, details of such postprocessing will not be fully revealed.

Overall, existing CD methods rely on numerous manually labeled data to optimize the model. Without the guidance of labels or only a handful of labels available, the performance of the model will be significantly degraded. Under this circumstance, the ability of the network structure is virtually weakened, and the model has a high probability of overfitting. *Can we explore the potential of the visual representation in a self-supervised strategy to alleviate the dependence on labels?* Motivated by this, we rethink SSL-based CD procedures. If we can acquire image pairs from different domains through adaptive methods, these interdomain samples could be identified as natural positive labels. As illustrated in Fig. 1, SSL pretraining can be first performed between the interdomain samples to get feature representations. Therefore, SSL pretraining can be first performed between the interdomain samples to get feature representations. After adequate labels are obtained, the pretrained representations will serve as guidance for downstreamed fully supervised fine-tuning to improve model performance.

In this article, we propose a self-supervised framework for bi-temporal CD on RS images via domain adaption. The SSLChange framework is organized in a two-stage approach, containing an interdomain adaptive encoder and a hierarchical contrastive head successively. The domain adapter (DA) adversarially gathers the distance between the original bi-temporal images and eliminates the effects of different imaging conditions. The hierarchical contrastive head is assigned to increase the similarity between features from unchanged regions. It is worth noting that since SSLChange adopts a SimSiam-like paradigm, the network only receives pseudo-positive pairs in latent space, which are generated by the DA from the same single-temporal image. The purpose of such an operation is to improve the robustness of pseudo-changes caused by data augmentation, focusing on the structural changes of ground objects at different imaging temporal.

The main contributions can be summarized as follows.

1) We propose a self-supervised framework named *SSLChange* for bi-temporal CD tasks on RS images that require no additional data or labels, which can be conveniently integrated into existing CD baselines.
2) We develop an *DA* module for bi-temporal SSL-based pretext tasks. In comparison to the original data augmentation, this transformation aligns more closely with real-world scenarios.
3) We design a *Hierarchical Contrastive Head* including spatial branch and channel branch to extract local and global features to effectively exploit pixel-level semantic information.

The rest of this article is organized as follows. Section II expounds previous work related to this article. The details of the proposed SSLChange framework are described in Section III. Extensive experimental results are presented in Section IV. Finally, the summary and perspectives are drawn in Section V.

II. Related Work

A. Domain Adaptation in RS Data

Interdomain variance is a common problem in multitemporal RS tasks [43], [44]. Since multitemporal images are acquired with different sensors and external environments, uncertain interdomain differences are introduced to the data [45]. Traditional manual interpretation is based on expert knowledge [46], [47]. However, existing statistical models may recognize such interdomain differences as different patterns. With increasing computational capability, convolutional neural networks (CNNs) are found to possess a certain degree of invariance to the semantic differences of multitemporal RS images and the extracted features are semantically

consistent [48], [49], [50]. However, when the interdomain differences are further expanded (e.g., dual-temporal images acquired in sunny and rainy seasons), the CNN fails to distinguish the interdomain variance on the images [51], [52], [53]. To tackle the interdomain variance, researchers have proposed the concept of domain adaptation, hoping to mine the mapping between latent spaces among the multitemporal data. Optimal latent features are expected to have unique correspondence on the original multitemporal images. Subsequent feature reconstruction is based on the invariant latent features, and the reconstructed multitemporal features share similar distributions.

The following two solutions to interdomain variance are generally adopted: distribution metric consistency-based approaches and adversarial-based approaches. An autoencoder is commonly employed in the former approaches, which performs feature reconstruction on embedding samples and measures the distance between the original samples and the reconstructed features to evaluate the performance [54]. Zhang et al. [55] proposed an unsupervised framework for multiresolution CD tasks. By the use of a denoising autoencoder, the interdomain variance between different images is largely eliminated. Sun et al. [56] introduced a coupled autoencoder into the bi-temporal HSI Image CD task, coping with the feature differences caused by spectral variation. However, the dimension of reconstructed features through an autoencoder is lower than the input samples, which brings uncertainties to the distance metric. Furthermore, the autoencoder is prone to "mode collapse" during training [57], [58]. Specifically, instead of learning the distribution of the input data, the model keeps outputting the same results as the input data. In this particular case, the loss of the model maintains a low value, but the performance cannot be satisfying [59].

The typical architecture for adversarial-based approaches is the generative adversarial network (GAN) [60]. The GAN performs feature extraction and reconstruction from random input noise vectors to generate fake samples similar to real ones. Meanwhile, the discriminator receives both real samples from the dataset and fake samples from the generator and then strives to judge the authenticity of the input data. The parameters of the GAN are optimized through iterative competition between the generator and the discriminator [61]. Chen et al. [62] proposed a CD model with adversarial augmentation to handle the data insufficiency. Li et al. [63] modified the vanilla GAN to translate optical and SAR images into the same feature domain and then performed CD between the translated heterogeneous data. However, the vanilla GAN may also suffer from the extreme case of mode collapse during training. Moreover, since the GAN generates objects from random noise vectors, the generated results are inevitably affected by random inputs, which is unacceptable in some pixel-level vision tasks [64], [65].

Overall, the above work demonstrates the benefit of domain adaptation methods in vision tasks, but the reconstructed features still need to be further refined to rule out wrong generation results. Therefore, it is necessary to construct a robust domain adaptation module to better serve downstream CD tasks.

B. Self-Supervised Contrastive Learning in the CD Task

Self-supervised representation learning has demonstrated success in NLP, but it still has great potential for visual tasks requiring pixel-level dense prediction. With the inspiration of self-supervision, visual tasks are undergoing a new development stage, of which contrastive self-supervised methods are an essential branch. Some contrastive self-supervised methods such as MoCo [24], SimCLR [25], BYOL [26], and SimSiam [27] have enabled models pretrained in label-free ways to outperform supervised pretraining models on many downstream tasks. In such contrastive methods, images are encoded into positive–negative pairs [24], [25] or double-positive pairs [26], [27], and then the Siamese architecture is utilized for feature extraction and comparison.

For the huge amount of data in the RS community, fine-grained labeling is a nearly impossible burden. The boom of contrastive self-supervised methods makes it possible to perform unlabeled pretraining in RS tasks, especially in RSCD tasks. Manas et al. [66] proposed a multiaugmentation contrastive self-supervised framework, named Seasonal Contrast (SeCo), which takes the effect of seasonal changes on feature representations into consideration to perform CD between Sentinel-2 satellite images. Jiang et al. [67] proposed a two-branch contrastive framework by embedding the CD backbone into a self-supervised paradigm to extract local and global features, which outperforms the supervised baseline. Muhtar et al. [68] proposed a contrastive mask image distillation framework in a teacher–student self-distillation architecture.

Although the aforementioned methods highly improve the efficiency and contribution of unlabeled data in RSCD tasks, random data augmentation is still required to construct pseudo-contrastive sample pairs when performing self-supervised pretraining [69], [70], [71], [72]. Uncertainty may be introduced during this process, causing fluctuations in model performance. In this regard, the proposed SSLChange framework takes the characteristics of multitemporal RSCD tasks into account, introducing domain adaptive into CD tasks to suppress the representation differences between multitemporal contrastive sample pairs to enhance the adaptability of interdomain tasks.

III. Proposed Method

In this section, we first elaborately introduce the overview of the proposed SSLChange framework. Then, the main components of SSLChange are described sequentially. The specific method of transferring the pretrained SSLChange framework to downstream RSCD tasks is also explained.

A. Overview

The overall workflow of SSLChange is illustrated in Fig. 2. Our approach builds upon the distillation-based SSL paradigm [24], [25], [27]. However, in contrast to the aforementioned methods, we originally propose two key components applicable to the RSCD task: *Domain Adapter* and *Hierarchical Contrastive Head*. The DA facilitates the transformation of bi-temporal patches from the CD dataset

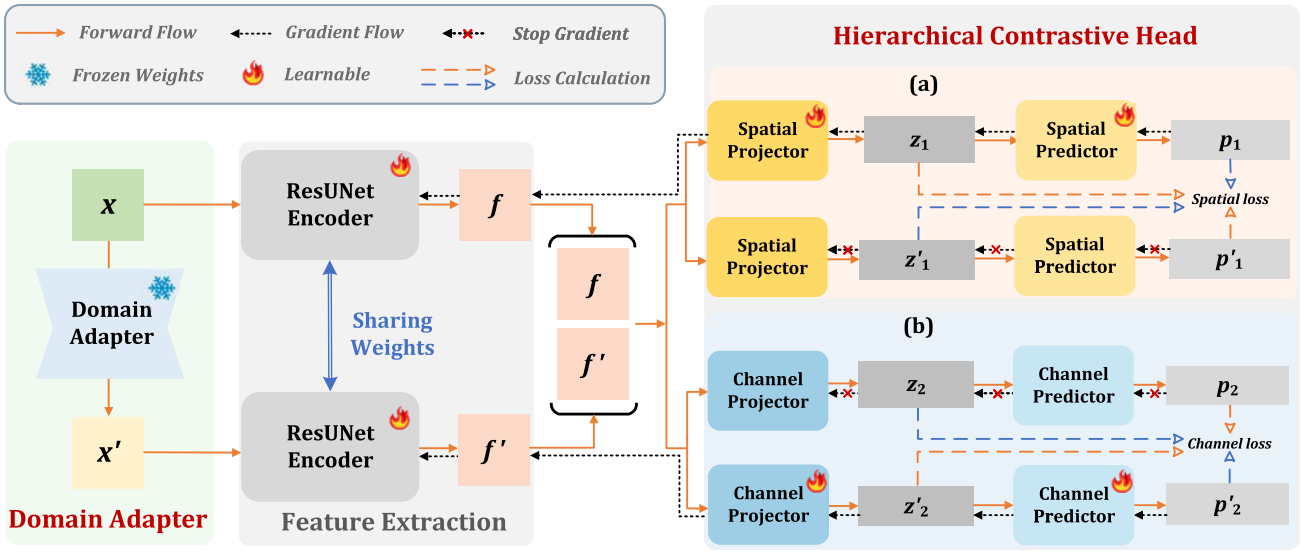

Fig. 2. Workflow of the proposed SSLChange framework for RSCD tasks. The whole SSLChange framework contains two main parts: DA and the hierarchical contrastive head. First, a transferred view x' is obtained through the DA from a single-temporal sample x. Then, the paired view $\{x, x'\}$ is processed by a ResUNet encoder with sharing weights to get feature set $\{f, f'\}$. The feature set is assigned to two branches: The spatial branch and the channel branch. Each branch consists of a projector and a predictor. The spatial branch captures geometric features by coupled convolutional units, while the channel branch distills the semantic feature by dimension reduction. (a) Spatial branch. (b) Channel branch.

into the latent space, subsequently reconstructing paired views. Following this, an encoder is used to extract features from paired images. Unless specifically stated, we employ ResUNet as the encoder. Finally, the Hierarchical Contrastive Head processes the paired features, performing feature contrast and similarity calculations within the Siamese branches. Next, we will provide a detailed introduction to these modules.

B. Domain Adapter

Many current SSL-based approaches [24], [25], [27] heavily depend on manual data augmentation to create diverse embedding views, necessitating manual tuning tailored to different pretext tasks. This often entails numerous trials involving various augmentation techniques to identify the optimal combination [25]. In the context of RSCD tasks, we revisit this process. Given that CD involves pairs of images captured at distinct time points, influenced by varying imaging conditions such as lighting and weather, we propose a different approach. Leveraging the variations in image style, we train a DA for bidirectional projection. This strategy enables us to generate a pair of naturally aligned positive samples.

To achieve this goal, the most straightforward way is to utilize existing image-to-image translation (I2IT) algorithms [73], [74], [75] to input images. Here, we illustrate using a GAN-based I2IT algorithm as a representative example. Given a pair of images $\{x_1, x_2\} \in \mathbb{R}^{H*W*C}$, where x_1 and x_2 share the same geolocation but differ in time phase and image styles, we establish a set of image translation and discriminative networks (G_1, D_1) and (G_2, D_2)

$$x_1' = G_1(x_1), \quad x_2' = G_2(x_2) \tag{1}$$

where x_1' and x_2' are the reconstructed samples. The discriminators D_1 and D_2 need to judge whether the input sample is generated or natural. G_1 and G_2 are trained adversarially to obtain the ability to project between the two image domains, which can generate positive samples. According to the analysis in our previous work [76], we choose modified CycleGAN [74] with stable performance as the DA.

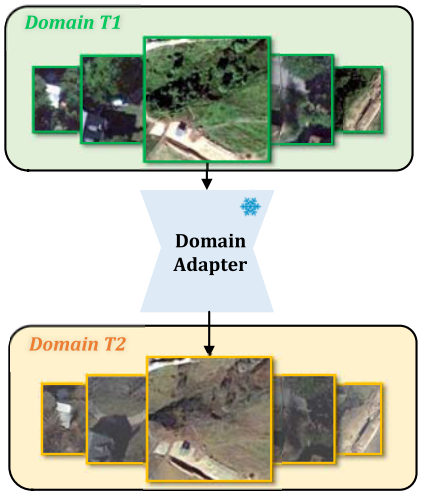

Fig. 3. Illustration of the DA for bi-temporal transformation on the binary CD dataset. The DA is used to project the samples from domain T_1 to the opposite domain T_2.

As shown in Fig. 3, we freeze the model parameters of G_1/G_2 as DA Θ_{ada}, which is used for the generation of transferred views. A single-temporal patch $x \in \mathbb{R}^{H*W*C}$ is sampled from the first temporal T_1 subset of the CD dataset. Its corresponding transferred view is $x' = \Theta_{\text{ada}}(x)$. View x' generated by Θ_{ada} differs from the source view x in image style, but shares the same content with view x. Therefore, we obtain paired views $\{x, x'\}$ from the generated results, which is naturally suitable for image contrastive learning. It is worth noting that in the specific implementation of SSLChange, only G_1 is employed as the DA. The experimental detail can be found in Section IV-D3.

C. Hierarchical Contrastive Head

As shown in Fig. 2, the generated paired views $\{x, x'\}$ will be fed into the encoder for feature extraction and Hierarchical Contrastive Head for further processing. The hierarchical contrastive head contains a spatial branch and a channel branch. Each branch contains a projector and a predictor.

1) Spatial and Channel Coding: In the spatial branch above, x and x' pass through the spatial projector and the predictor, respectively, to get intermediate code $\{z, z'\}$ and deep code $\{p, p'\}$. The specific calculation is as follows:

$$\begin{cases} z_1 = \textbf{Spa\_Proj}(\mathbf{E}(x)), & p_1 = \textbf{Spa\_Pred}(z_1) \\ z_1' = \textbf{Spa\_Proj}(\mathbf{E}(x')), & p_1' = \textbf{Spa\_Pred}(z_1') \end{cases} \tag{2}$$

where $\mathbf{E}$ is the ResUNet encoder. Similarly, x and x' pass through the channel projector and the predictor, respectively, in the below branch to obtain $\{z_2, z_2'\}$ and $\{p_2, p_2'\}$

$$\begin{cases} z_2 = \textbf{Cha\_Proj}(\mathbf{E}(x)), & p_2 = \textbf{Cha\_Pred}(z_2) \\ z_2' = \textbf{Cha\_Proj}(\mathbf{E}(x')), & p_2' = \textbf{Cha\_Pred}(z_2'). \end{cases} \tag{3}$$

Architecture Settings: The specific architecture of the components in the SSLChange framework is as follows.

1) *ResUNet Encoder:* The ResUNet encoder uses the U-Net structure with a ResNet-18 backbone. The size of the output feature is consistent with the input. The features from layers [2, 4, 5, 6, 7] are extracted and unsampled for fusion.
2) *Spatial Projector & Predictor:* The spatial projector and the predictor share the same structure, both consisting of two basic convolutional units with BN and ReLU applied to them. The size of the output feature is consistent with the input.
3) *Channel Projector & Predictor:* The channel projector consists of a three-layer MLP with a BN layer. The output layer has no BN or activation layer. It is a two-layer MLP. The former layer applies BN. The output of both modules is a 2048-D vector.

2) Hierarchical Contrastive Learning: We employ spatial and channel dimension operations to dig the deep structure and semantic features in RS images, and then feature contrast is performed based on the aforementioned features. It is worth noting that we do not introduce any additional labels in this step, only a hierarchical cross-contrastive mechanism is established between the pipelines in each branch to constrain the network optimization. This prompts the model to decrease its reliance on labels by the self-supervised contrastive mechanism, further exploiting the spatial and semantic representational capabilities of the model.

To achieve this, the feature distance between pipelines in each branch needs to be pulled closer. Denoting the two output codes as z_1 and p_1', we minimize the negative cosine similarity between two codes

$$D(z_1, p_1') = -\left\langle \frac{z_1}{\|z_1\|_2}, \frac{p_1'}{\|p_1'\|_2} \right\rangle. \tag{4}$$

Specifically, we obtain the spatial codes sets z_1, p_1 and z_1', p_1' through previous calculations. The calculation for loss in the spatial branch is

$$\mathcal{L}_{\text{Spa}} = \frac{D(z_1, p_1')}{2} + \frac{D(z_1', p_1)}{2}. \tag{5}$$

Similarly, the loss in the channel branch is calculated by

$$\mathcal{L}_{\text{Cha}} = \frac{D(z_2, p_2')}{2} + \frac{D(z_2', p_2)}{2}. \tag{6}$$

The total loss is the linearly weighted sum of the spatial contrastive loss and the channel contrastive loss

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{Spa}} + (1 - \alpha) \cdot \mathcal{L}_{\text{Cha}} \tag{7}$$

where α represents the weight parameter of the spatial branch.

Through the aforementioned optimization, the model achieves a balance between the spatial coding and semantic coding branches and possesses the capability to extract image structural and semantic features more accurately. It should be noted that, during the optimization process in each branch, the stop-gradient operation is performed on one of the two pipelines, and the gradient will only be backpropagated along the other pipeline. This operation ensures that the model avoids collapsing in the absence of negative samples. The experimental detail can be found in Section IV-D3.

To facilitate a better comprehension of our approach, we present the pseudo-code of the proposed SSLChange framework in Algorithm 1.

Algorithm 1 SSLChange PseudoCode, PyTorch-Like

```
# DA: Domain Adapter
# E: ResUNet Encoder
# Spa_Proj/Pred: Spatial projector/predictor
# Cha_Proj/Pred: Channel projector/predictor

for x in loader:
   x' = DA(x) # generate transferred view x'
   f, f' = E(x), E(x') # extract features

   # Spatial projections& predictions
   z1, z1' = Spa_Proj(f), Spa_Proj(f')
   p1, p1' = Spa_Pred(z1), Spa_Pred(z1')

   # Channel projections& predictions
   z2, z2' = Cha_Proj(f), Cha_Proj(f')
   p2, p2' = Cha_Pred(z2), Cha_Pred(z2')

   # calculate losses
   L_Spa = D(z1, p1')/2 + D(z1', p1)/2
   L_Cha = D(z2, p2')/2 + D(z2', p2)/2
   L = α· L_Spa + (1−α)·L_Cha

   L.backward()   # back-propagate
   update(E)   # update parameters
   update(Spa_Proj, Spa_Pred)
   update(Cha_Proj, Cha_Pred)

def D(p, z) # negative cosine similarity
   z = z.detach() # stop gradient

   p = normalize(p, dim=1) # L2-norm on p
   z = normalize(z, dim=1) # L2-norm on z
   return -(p*z).sum(dim=1).mean()
```

D. Downstream Fine-Tuning

In this section, we delve into the process of implementing the SSL-based pretrained model for downstream pixel-level tasks, addressing a crucial and often overlooked detail.

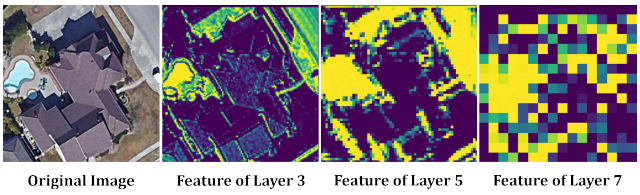

Fig. 4. Visualization of pretrained features from different layers.

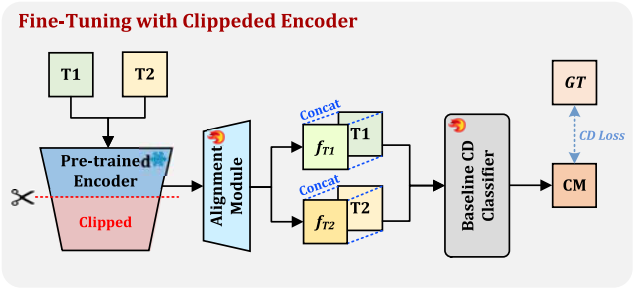

Fig. 5. Illustration of the fine-tuning method with a clipped pretrained encoder.

While existing works predominantly focus on applying SSL-based pretraining to object-level or instance-level tasks like image classification or object detection, there is a significant gap in understanding how to adapt this approach to pixel-level tasks, specifically CD. In the former cases, SSL-based pretraining focuses on the image-level tasks where the output of the pretrained encoder serves as the input to the bottleneck layer during transfer to downstream tasks. However, the scenario becomes more intricate when dealing with pixel-level tasks such as CD. In this context, the baseline expects an input image with dimensions of (C, H, W), which does not readily align with the output vector dimension (e.g., 2048) generated by the pretrained encoder.

We first attempt to directly freeze the pretrained encoder as a feature extractor and connect it to the CD baseline. However, we find that the outputs of the pretrained encoder usually contain high-dimensional semantic features, which cannot intuitively reflect the target distribution as the original image. In addition, such rigid migration may lead to gradient vanishing and performance degradation in CD tasks.

For a clear presentation, we visualize the pretrained features from the frozen encoder in Fig. 4. Considering that the shallow features contain more structural features such as edges and corners, we clip the pretrained encoder and only extract the first three shallow features of the output. To solve the problem of dimensional unalignment between the output features and the original image, we find that only a lightweight alignment module containing two deconvolutional layers is sufficient to maintain the model performance to the utmost degree.

As shown in Fig. 5, given a pair of images $\{T1, T2\}$ to be detected, we use the clipped pretrained encoder to extract shallow features and restore the feature shape through the alignment module. Then, the aligned features are concatenated with the original images in the channel dimension as the embedding of the downstream baseline network. Finally, the change map (CM) calculated by the baseline network is compared with the (GT) to fine-tune and optimize using supervision. The optimization method is as follows:

$$\mathrm{CM} = \mathcal{F}(\mathrm{Concat}(T_1, f_{T1}), \mathrm{Concat}(T_2, f_{T2})) \tag{8}$$

$$\textbf{min} \quad \mathcal{L}(\mathrm{CM}, \mathrm{GT}) \tag{9}$$

where CM is the change map calculated by the baseline and GT is the ground truth. $\mathcal{F}$ is the downstream baseline and $\mathcal{L}$ is the loss function corresponding to the downstream baseline. In the case of SNUNet [77], we denote the detailed calculation of the loss function as follows:

$$\begin{cases} \mathcal{L}(y, \hat{y}) = \mathcal{L}_{\mathrm{WCE}} + \mathcal{L}_{\mathrm{Dice}} \\ \mathcal{L}_{\mathrm{WCE}} = -\omega \cdot \hat{y} \cdot \log(y) - (1 - \hat{y}) \cdot (\log(1 - y)) \\ \mathcal{L}_{\mathrm{Dice}} = 1 - \dfrac{2 \cdot \hat{y} \cdot \mathrm{softmax}(y)}{\hat{y} + \mathrm{softmax}(y)} \end{cases} \tag{10}$$

where ω is the weight parameter of the positive samples. y and $\hat{y}$ are the CM and GT, respectively.

The experimental detail of downstream fine-tuning can be found in Sections IV-D1 and IV-D2.

TABLE I
SETTINGS OF THE LEVIR-CD SERIES DATASETS. (THE NUMBERS IN THE TABLE ARE A PAIR OF IMAGES)

Subset	Resolution	Train	Validation	Test
LEVIR-CD-10%	0.5 m/pixel	45	13	64
LEVIR-CD-20%		89	26	
LEVIR-CD-50%		223	64	
LEVIR-CD		445	128	

IV. EXPERIMENTS

A. Datasets

In this section, we introduce the datasets used in the experiments. It is worth noting that we want to verify the effectiveness of the proposed SSLChange framework in the extreme case of limited data, so each selected dataset is randomly diluted by the ratio of 10%, 20%, and 50%. Comparative experiments are performed both on six diluted datasets and two original datasets simultaneously.

1) LEVIR-CD Series Datasets: The LEVIR-CD dataset [78] contains 673 VHR bi-temporal RS image pairs, and the resolution of each image is 0.5 m. The original size of the image is 1024×1024 pixels and the entire dataset is split by the ratio of 7:2:1 into training/validation/test sets. To ensure consistent testing standards, the number of test sets was not diluted. The main labeling object is the changes in buildings between two time phases. The specific settings of the LEVIR-CD series datasets are shown in Table I.

2) CDD Series Datasets: The CDD dataset [79] contains 16 000 pairs of VHR RS images, with resolutions ranging from 0.03 to 1 m. The size of the original image is 256*256 pixels. The provider of the CDD dataset splits the entire dataset into 10 000/3000/3000, so we keep this split in the experiments. The main labeling objects are changes in buildings, roads, and vehicles. The specific settings of the CDD series datasets are shown in Table II.

Note that we use the entire unlabeled LEVIR-CD and CDD datasets during the SSLChange pretraining. Then, the diluted

TABLE II
SETTINGS OF THE CDD SERIES DATASETS. (THE NUMBERS IN THE TABLE ARE A PAIR OF IMAGES)

Subset	Resolution	Train	Validation	Test
CDD-10%	0.03m-1 m/pixel	1000	300	3000
CDD-20%		2000	600	
CDD-50%		5000	1500	
CDD		10000	3000	

datasets are used for downstream fine-tuning to evaluate the effectiveness of the pretrained encoder from the SSLChange framework under limited data.

B. Evaluation Metrics

In evaluating the performances of the SSLChange framework, we select four evaluation metrics: Precision, Recall, $F1$-Score, and IoU. The specific calculation method for metrics is as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\ F1 &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{IoU} &= \frac{\text{TP}}{\text{TP}+\text{FN}+\text{FP}} \end{aligned} \tag{11}$$

where TP, FP, and FN are components in the confusion matrix, representing true positives, false positives, and false negatives, respectively.

C. Implementation Details

1) Selected Downstream RSCD Baselines: We select five open-source RSCD baselines to evaluate the effectiveness of the proposed SSLChange framework transferred to them.

1) *FC-EF [80]:* Fully convolutional network with a single-branch U-Net architecture. Bi-temporal images are fused early before entering the network. Skip connection is used in the upsampling and downsampling stages.
2) *FC-Siam-conc [80]:* Dual-branch Siamese architecture with a full convolutional network. Each branch receives an image of a single phase and performs feature concatenation fusion and long-range skip connection during the upsampling stage.
3) *FC-Siam-diff [80]:* Dual-branch Siamese architecture with a full convolutional network. The difference map is calculated during the downsampling stage. Feature concatenation fusion and long-range skip connection are performed during the upsampling stage.
4) *SNUNet-CD [77]:* Dual-branch U-Net++ architecture. Hierarchical features of the input images are connected densely, and the ECAM attention mechanism is performed to focus on the features.
5) *USSFCNet [81]:* A lightweight architecture design is adopted. Multiscale decoupled convolution and spatial–spectral feature cooperation strategies are introduced to extract richer features.

2) Model Training and Testing: The training and inference processes are implemented on an NVIDIA RTX 3090 GPU with 24 GB of memory. The SSLChange framework is trained for 100 epochs with a batch size of 8. The parameters are optimized by an SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate is set to 0.001 with cosine decay. In the downstream fine-tuning, the selected baselines are optimized by the AdamW optimizer with a momentum of 0.999 and a weight decay of 0.01. Considering the limited data volume in downstream tasks, the batch size for fine-tuning is set to 4. Furthermore, we discard all random operations in the augmentation during downstream fine-tuning to eliminate the influence of randomness, and only format transformation functions are retained.

D. Experimental Results

1) Quantitative Analysis: In this part, we perform comparative experiments on the selected CD baselines on two entire datasets and six diluted datasets. Specifically, we evaluate the improvement of the proposed SSLChange framework over existing CNN-based CD baselines, especially the performance and stability in data-limited situations.

a) CDD series datasets: The specific results of comparative experiments on the CDD series datasets are shown in Table III. The results reflect that the proposed SSLChange framework provides a large gain for the existing RSCD baselines on the CDD series datasets. The green subscripts in the table represent the performance gain of the baselines with the SSLChange framework applied compared to the original baselines. On the CDD series datasets, SSLChange showed a positive impact on most of the baselines, with the largest gains in the main evaluation metrics $F1$ and IoU reaching 15.36% and 15.15%, respectively. In terms of precision and recall metrics, SSLChange also showed a great improvement, which shows that the SSLChange framework helps the baselines more accurately and comprehensively detect the real change regions between the bi-temporal images. From the horizontal comparison, it can be found that when the datasets are diluted to 10% or 20%, the performance of the original baselines shows a large degradation. During the training, we also found that after removing the random augmentation in the original baseline, the performance of the baselines would become oscillatory, and the results of several training sessions under the same configuration showed drastic fluctuations. While the application of the SSLChange framework can help the baselines return to a relatively stable performance. We notice that there exist two outliers in the result table, respectively, in FC-EF (on the CDD-50% dataset) and SNUNet-CD (on the CDD dataset). After analysis, we consider that the fine-tuning strategy applied is to freeze the parameters of the clipped pretrained encoder for feature extraction, and then an upsampling alignment module is added to the baseline. This operation increases the number of parameters and complexity of the baselines. In addition, the CDD dataset has a relatively large amount of data, so it fails to show advantages under limited training epochs.

b) LEVIR-CD series datasets: The specific results of comparative experiments on the LEVIR-CD series datasets

TABLE III

COMPARISON RESULTS ON DATASETS WITH DIFFERENT SAMPLING RATIOS FROM THE CDD DATASET. THE GREEN SUBSCRIPTS IN THE TABLE REPRESENT THE PERFORMANCE GAIN OF THE BASELINES WITH THE SSLCHANGE FRAMEWORK APPLIED COMPARED TO THE ORIGINAL BASELINES

Baseline	+SSLChange	CDD-10%				CDD-20%				CDD-50%				CDD			
		Precision	Recall	**F1**	**IoU**	Precision	Recall	**F1**	**IoU**	Precision	Recall	**F1**	**IoU**	Precision	Recall	**F1**	**IoU**
FC-EF [80]	✗	69.62	64.91	67.18	50.58	89.51	62.72	73.76	58.43	90.52	77.04	83.24	71.29	95.23	79.39	86.59	76.35
	✓	77.10	60.37	67.72	51.19	81.99	70.40	75.75	60.97	86.03	76.52	81.01	68.07	91.43	83.83	87.47	77.72
	Δ			+0.54	+0.61			+1.99	+2.54			-2.23	-3.22			+0.88	+1.37
FC-Siam-conc [80]	✗	81.63	38.51	52.34	35.44	83.23	44.06	57.62	40.46	88.96	48.19	62.51	45.47	94.62	34.21	50.25	33.56
	✓	79.16	40.08	53.22	36.26	79.82	46.15	58.49	41.33	85.47	52.27	64.87	48.01	90.01	51.50	65.51	48.71
	Δ			+0.88	+0.82			+0.87	+0.87			+2.36	+2.54			+15.36	+15.15
FC-Siam-diff [80]	✗	85.98	30.81	45.36	29.33	88.72	37.96	53.17	36.21	92.15	37.63	53.44	36.46	93.26	36.35	52.31	35.42
	✓	83.05	35.32	49.57	32.95	86.88	39.21	54.03	37.02	88.59	43.09	57.98	40.82	90.91	47.47	62.37	45.32
	Δ			+4.21	+3.62			+0.86	+0.81			+4.54	+4.36			+10.06	+9.90
SNUNet [77]	✗	71.69	69.80	70.73	54.71	91.83	60.14	72.68	57.09	91.32	74.46	82.03	69.54	92.51	83.99	88.05	78.64
	✓	96.76	65.57	74.69	59.60	89.37	73.16	80.46	67.30	90.97	79.98	85.12	74.10	90.10	79.12	84.26	72.80
	Δ			+3.96	+4.89			+7.78	+10.21			+3.09	+4.56			-3.79	-5.84
USSFCNet [81]	✗	84.94	57.34	68.46	52.05	87.73	72.70	79.51	65.99	89.58	80.99	85.07	74.01	90.79	91.40	91.10	83.65
	✓	81.01	64.56	71.86	56.07	83.56	77.21	80.26	67.03	90.41	84.44	87.32	77.49	92.82	90.03	91.40	84.17
	Δ			+3.40	+4.02			+0.75	+1.04			+2.25	+3.48			+0.30	+0.52

TABLE IV

COMPARISON RESULTS ON DATASETS WITH DIFFERENT SAMPLING RATIOS FROM THE LEVIR-CD DATASET. THE GREEN SUBSCRIPTS IN THE TABLE REPRESENT THE PERFORMANCE GAIN OF THE BASELINES WITH THE SSLCHANGE FRAMEWORK APPLIED COMPARED TO THE ORIGINAL BASELINES

Baseline	+SSLChange	LEVIR-10%				LEVIR-20%				LEVIR-50%				LEVIR			
		Precision	Recall	**F1**	**IoU**	Precision	Recall	**F1**	**IoU**	Precision	Recall	**F1**	**IoU**	Precision	Recall	**F1**	**IoU**
FC-EF [80]	✗	68.58	58.96	63.40	46.41	78.01	73.42	75.65	60.83	82.87	74.13	78.26	64.28	87.11	74.01	80.03	66.71
	✓	67.11	61.66	64.27	47.35	76.87	71.73	74.21	59.01	81.15	76.49	78.75	64.95	83.20	79.77	81.45	68.70
	Δ			+0.87	+0.94			-1.24	-1.82			+0.49	+0.67			+1.42	+1.99
FC-Siam-conc [80]	✗	76.76	58.46	66.68	50.01	80.41	65.96	72.47	56.83	83.71	71.93	77.37	63.09	84.03	72.68	77.94	63.86
	✓	68.99	68.07	68.53	52.12	78.21	67.12	73.85	58.54	82.72	77.01	79.76	66.34	83.53	77.62	80.47	67.32
	Δ			+1.85	+2.11			+1.38	+1.71			+2.39	+3.25			+2.53	+3.46
FC-Siam-diff [80]	✗	76.64	28.50	41.55	26.22	83.21	61.02	70.41	54.33	84.23	64.33	72.95	57.42	84.54	72.22	77.89	63.79
	✓	74.61	54.25	62.83	45.80	81.34	63.18	71.12	55.19	85.52	65.64	74.27	59.07	84.88	73.34	78.69	64.86
	Δ			+21.31	+19.58			+0.71	+0.86			+1.32	+1.65			+0.80	+1.07
SNUNet [77]	✗	65.79	58.96	62.19	45.12	68.70	59.79	63.94	46.99	83.78	64.52	72.90	57.35	78.89	67.71	72.88	57.32
	✓	64.93	62.94	63.92	46.97	70.97	61.11	65.68	48.89	82.46	67.91	75.18	60.23	84.76	76.19	80.25	67.01
	Δ			+1.73	+1.85			+1.74	+1.90			+2.28	+2.88			+7.40	+9.69
USSFCNet [81]	✗	69.94	69.42	66.97	53.46	76.75	75.56	76.15	61.49	82.76	80.07	81.39	68.62	82.01	85.54	83.74	72.02
	✓	72.60	70.54	71.56	55.71	78.91	76.89	77.89	63.78	83.51	80.09	81.76	69.15	84.25	83.36	83.81	72.13
	Δ			+4.59	+2.25			+1.74	+2.29			+0.37	+0.53			+0.07	+0.11

are shown in Table IV. The proposed SSLChange framework also shows an improvement effect on the LEVIR-CD series datasets. It is worth noting that the data volume in the LEVIR-CD series datasets is smaller than that of the CDD series datasets. Therefore, the data limitation in the LEVIR-CD series datasets is more serious. In this case, some of the selected baselines also show unstable and overfitting. When the amount of data is at 50% or the full ratio, the SSLChange framework can provide satisfied gains for most baselines. In the extreme cases where the dilution ratio decreases to 10% and 20%, the SSLChange framework can help the performance of the baselines return to a relatively stable and homogeneous level. The maximum improvement of the $F1$ and IoU metrics reaches 21.31% and 19.58%, which proves the performance advantage of the SSLChange framework in extreme cases with limited data. In addition, we observe an outlier in FC-EF (on the LEVIR-CD-20% dataset). After a comprehensive analysis, we conclude that FC-EF, as a lightweight network with a simple structure, is more prone to overfitting in extreme cases with limited data. By comparison, we find that in the LEVIR-CD-20% dataset, FC-EF presents a falsely high index. The improvement of the SSLChange framework under the small amount of data pursues stability and homogeneity, which is lower than the individual metric of FC-EF.

On both the CDD series datasets and LEVIR-CD series datasets, we observe that in some cases, the precision of

TABLE V

VISUAL COMPARISON OF BASELINES AND OUR PROPOSED SSLCHANGE METHOD ON THE CDD SERIES DATASETS

Method	Testing Samples			CDD-10%		CDD-20%		CDD-50%		CDD	
	T1	T2	GT	Sup	SSLChange	Sup	SSLChange	Sup	SSLChange	Sup	SSLChange
FC-EF [80]											
FC-Siam-conc [80]											
FC-Siam-diff [80]											
SNUNet [77]											
USSFCNet [81]											

TABLE VI

VISUAL COMPARISON OF BASELINES AND OUR PROPOSED SSLCHANGE METHOD ON THE LEVIR-CD SERIES DATASETS

Method	Testing Samples			LEVIR-10%		LEVIR-20%		LEVIR-50%		LEVIR	
	T1	T2	GT	Sup	SSLChange	Sup	SSLChange	Sup	SSLChange	Sup	SSLChange
FC-EF [80]											
FC-Siam-conc [80]											
FC-Siam-diff [80]											
SNUNet [77]											
USSFCNet [81]											

the proposed SSLChange framework is slightly lower than the baselines. But for the remaining metrics, SSLChange outperforms significantly the baselines. We believe that the reason is that with limited training samples, the baseline is usually overfitted, leading to falsely high precision. Aligned with the SSLChange framework, the baselines exhibit a more generalized recognition capability, allowing the model to achieve higher recall metrics at the acceptable expense of accuracy. As a consequence, more regions of real change will be correctly recognized.

In addition, we find a gap in the trend of increase and decrease in precision and recall between the selected baselines on different datasets. Overall, the precision and recall of almost all baselines change asynchronously on the CDD series datasets. While on the LEVIR-CD series datasets, precision and recall of SNUNet and USSFCNet increase synchronously. Such performance variation is related to the utilization of the attention mechanisms. SNUNet and USSFCNet employ ECAM and SSFC attention, respectively. This helps the models to focus on local details in high-resolution images, which is reflected in an increase in precision metric. Moreover, benefiting from the constraint from the SSLChange on the global feature, the recall metrics on LEVIR-CD series datasets increase simultaneously. However, the samples in

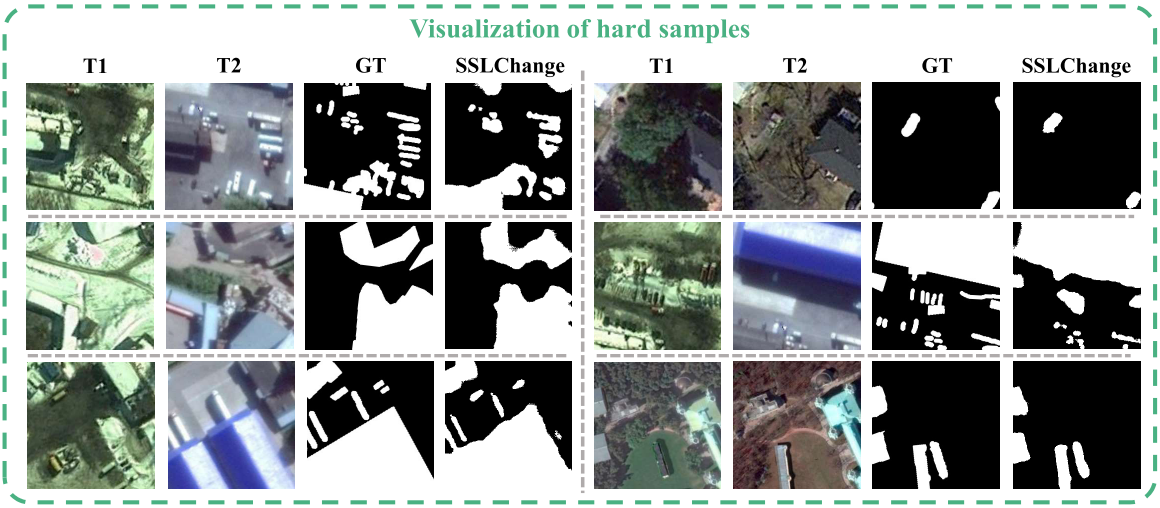

Fig. 6. Performance of the proposed SSLChange in hard samples with complex scenarios.

the CDD series dataset are generally of low to medium resolution. The attention mechanism is limited by the image quality to accurately capture local information, leading to a decrease in precision metric. Extensive experimental results show that the SSLChange framework motivates the model to maintain global feature extraction regardless of the quality of the dataset. In realistic scenarios, recall metric is usually given higher priority than precision to ensure that as many changed regions as possible are accurately detected. In addition, considering that $F1$ is a more comprehensive metric than precision and recall, our proposed SSLChange still obtains the highest performance on $F1$ despite the decrease in precision.

2) Visualization Analysis: To show the improvement effect of the SSLChange framework more intuitively, we perform visualizations on two entire datasets and six diluted datasets. The visualization results of the CDD and the LEVIR-CD series datasets are listed in Tables V and VI. The main targets in the CDD series datasets are scattered single buildings and linear roads, while the main targets in the LEVIR-CD series datasets are dense clusters of buildings. Through comparison, we find that the SSLChange framework improves the capability of geometric structure feature acquisition for CD baselines under both data distributions. In the case of limited data, SSLChange can obtain more satisfied feature segmentation results than the original baselines. We believe that the proposed SSLChange framework brings benefits to the baseline for two reasons. The application of the spatial projector and the predictor in the SSLChange framework helps to maintain and capture accurate spatial features during the pretext tasks. In addition, the channel branch in SSLChange reduces the dimension of the samples by the MLP module, which makes the framework more sensitive to interclass differences, leading to more accurate segmentation results in the latent space.

For better illustration, we additionally visualize several hard samples with complex scenarios. As illustrated in Fig. 6, although the presented samples with significant background change and shadow, our proposed SSLChange is still able to accurately capture the changed regions.

TABLE VII
ABLATION EXPERIMENTS OF DA ON THE LEVIR-CD-50% DATASET

Baseline	Method	F1	IoU
SSLChange	w/o Domain Adapter	74.23	59.02
	w/ Domain Adapter	75.18	60.23

3) Ablation Experiments: In this part, we perform several ablation experiments to evaluate the main components of the proposed SSLChange framework. The ablation study contains five main experiments.

a) Domain adapter: To evaluate the performance of the DA, the SSLChange framework is first trained without DA and with DA, respectively. Then, the pretrained encoder is clipped and aligned with the baseline to perform downstream fine-tuning. Specifically, in the case where the framework does not apply DA, the input samples x follow the random data augmentation method in SimSiam [27] and are sent into the Hierarchical Contrastive Head. We select SNUNet-CD as the baseline for comparison on the LEVIR-CD-50% dataset. The results of the ablation study are shown in Table VII. *w/o DA* represents without the Doamin Adapter, and *w/ DA* represents with the Doamin Adapter. The performance of $F1$ and IoU metrics demonstrates the superiority of the proposed DA. The adaptive translation between bi-temporal image domains promotes the encoder to capture cross-domain features and eliminate the effects of imaging conditions.

b) Encoder clipping: In the SSLChange framework, the pretrained encoder is frozen and clipped, and only the first three layers are preserved. Likewise, we explore the effect of pretrained encoder clipping. From the results shown in Table VIII, we observe a significant improvement in the performance of the SSLChange framework with encoder clipping, which corroborates the feature visualization results in Fig. 4. It demonstrates that shallow pretrained features from pixel-level SSL pretext tasks are more instructive for downstream fine-tuning.

c) Hierarchical contrastive head: The ablation study on the spatial projector and the predictor and the channel projector

TABLE VIII
ABLATION EXPERIMENTS OF PRETRAINED ENCODER CLIPPING ON THE LEVIR-50% DATASET

Baseline	Method	F1	IoU
SSLChange	w/o Encoder Clipping	72.86	57.31
	w/ Encoder Clipping	75.18	60.23

TABLE IX
ABLATION EXPERIMENTS OF THE CHANNEL AND SPATIAL PROJECTOR AND THE PREDICTOR ON THE LEVIR-CD-50% DATASET

Baseline	+ Spa.	+ Cha.	F1	IoU
SNUNet-CD	✗	✗	72.90	57.35
	✗	✓	73.49	58.09
	✓	✗	74.10	58.86
	✓	✓	75.18	60.23

and the predictor in the Hierarchical Contrastive Head is conducted in this section. Similarly, we select SNUNet-CD as the baseline for comparison on the LEVIR-CD-50% dataset. The results of the ablation study are shown in Table IX, where *Spa.* represents the spatial projector and the predictor and *Cha.* is the channel projector and the predictor.

It is worth noting that in the case where only the channel branch is applied, the method degenerates to SimSiam [27] as a comparison. The applications of both spatial and channel modules in SSLChange pretraining are proved to provide considerable benefits to the baseline. The spatial branch performs identical convolution operations to maintain the dimension of the features, while the channel branch obtains semantic information by feature reduction.

d) Downstream fusion: For the downstream fine-tuning in the SSLChange framework, the concatenation operation is applied to fuse the original embedding and the pretrained features. Here, we further exploit the impact of several fusion operations (i.e., add, multiply, deconvolution, and concatenation) for the downstream fine-tuning, without significantly increasing the whole model parameters.

The experimental results demonstrate that the concatenation fusion utilized in SSLChange shows significant superiority over the multiply ($F1$: 71.60%, IoU: 55.76%) and the deconvolution ($F1$: 65.58%, IoU: 48.78%) methods, obtaining a performance of 75.18% and 60.23% on $F1$ and IoU metrics, respectively. When compared to the add method, the concatenation method possesses a slight advantage ($F1$: 75.13%, IoU: 60.17%). We argue that the add fusion assigns equal weights to the two components for fusion, and the concatenation fusion preserves all feature components and allocates the weights automatically through the subsequent linear layer. Therefore, we select the concatenation fusion in our implementation.

e) Weight parameter α for loss function: The result curves on the impact of parameter α is presented in Fig. 7. The changing trend reflects that the proposed method achieves the optimal performance when the spatial loss and channel loss reach a relative balance ($\alpha = 0.5$). It could be concluded that when the local details from the spatial branch are in an equilibrium constraint with the global semantic codes from the channel branch, the encoder and the contrastive head are motivated to accurately capture valuable feature representations.

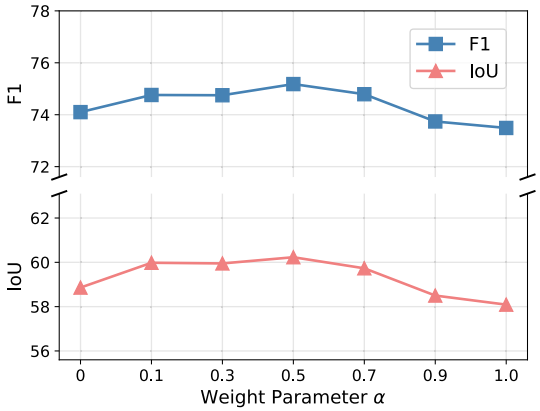

Fig. 7. Performance of $F1$ and IoU under different weight parameter α.

TABLE X
CROSS-DATASET VALIDATION FROM A LARGE-SCALE PRETRAINING DATASET TO A SMALL-SCALE DOWNSTREAM DATASET

Pre. Paradigm	LEVIR-10%		LEVIR-20%		LEVIR-50%		LEVIR	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Un. LEVIR-CD	63.92	46.97	65.68	48.89	75.18	60.23	80.25	67.01
Un. CDD	59.71	42.56	69.26	52.97	76.31	61.70	79.58	66.09

4) Cross-Dataset Validation: To evaluate the generalization ability of the proposed SSLChange framework, we perform cross-dataset validation between the two selected datasets. Considering that only a few labeled data are available in practical scenarios, we transfer the feature representations from a large-scale pretraining dataset to a small-scale downstream dataset. Specifically, we select the CDD dataset as the benchmark for the pretraining of SSLChange. Then, the downstream CD baseline is fine-tuned with the LEVIR-CD series datasets. The experimental results are shown in Table X, where *Un. LEVIR-CD* and *Un. CDD* represent SSLChange pretraining with unlabeled LEVIR-CD and unlabeled CDD datasets, respectively. Despite the large differences between the downstream dataset (LEVIR-CD series) and the pretrained CDD dataset, transferring the CDD pretrained model to the LEVIR-CD dataset still achieves satisfying performance, which is comparable to the LEVIR-CD pretraining results. Therefore, the generalization capability of the proposed SSLChange framework is further validated through the cross-dataset validation experiments.

5) Computational Efficiency and Stability: Computational efficiency and stability are also essential attributes of the proposed SSLChange framework.

a) Pretraining time cost of SSLChange: To get a tradeoff between data volume and model performance, we select a mini-batch of 8 to conduct pretraining with the SSLChange framework. We record the pretraining time cost for 100 epochs on two entire datasets: CDD and LEVIR-CD datasets in Table XI. The proposed SSLChange framework can accomplish pretraining with high efficiency regardless of the scale of the dataset. We observe that the SSLChange framework presents a higher pretraining speed on the LEVID-CD dataset

TABLE XI
PRETRAINING TIME COST OF THE PROPOSED SSLCHANGE FRAMEWORK

Dataset	Pre. Batch Size	Data Volume	Total Time	Speed
CDD	8	10000	6.25 h	44 FPS
LEVIR-CD		445	13.3 min	56 FPS

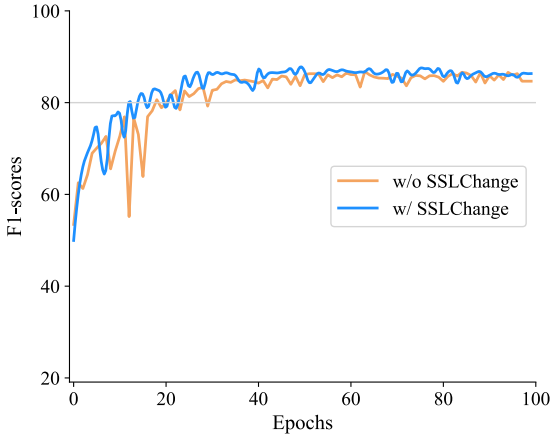

Fig. 8. Computational stability and convergence speed of the proposed method during the training process as $F1$-Epoch curves.

than that on the CDD dataset, which is most probably related to limited computing resources.

b) Downstream fine-tuning stability: From the curves shown in Fig. 8, it can be clearly observed that, in the downstream fine-tuning, the training process of the baseline with SSLChange applied is more stable within a limited number of epochs. From the amplitude of the curves, it can also be discovered that SSLChange promotes the baseline to converge faster to the optimal value.

E. Limitation Analysis

In the SSLChange framework, a DA and a hierarchical contrastive head are utilized to improve the performance of CD between bi-temporal optical images with self-supervised contrastive pretraining. The experimental results demonstrate the ability to stably enhance the CD baseline in a data-limited situation. However, several existing limitations as follows require further exploration, which also enlightens the future research directions.

1) Pretraining Paradigm: Self-supervised contrastive paradigm is adopted in SSLChange which motivates to obtain high-level semantic features. In contrast, the self-supervised mask modeling paradigm can promote the model to capture low-level local features by random mask reconstruction. Thus, the integration of the hierarchical feature extraction capabilities from different paradigms in CD tasks is required to be explored.

2) Multisource Domain Adaption: The SSLChange framework focuses on the domain adaption between the same source data. However, current airborne and spaceborne platforms are commonly equipped with multisensors to jointly acquire data for complementary purposes. Therefore, the generalized domain adaption between multisource data (e.g., SAR, LiDAR, and Text information) remains to be investigated to handle the various data emerging in the RS field.

3) Robustness to Pseudo-Changes: Some environmental factors and noises could disturb the model performance, such as seasonal background variations, label noise, unalignment noise, and object occlusion. These factors are also expected to be considered in future research.

V. CONCLUSION

In this article, we propose an SSLChange framework for bi-temporal CD tasks, which can be flexibly adapted to existing downstream CD baselines with a concise structure. Specifically, the transferred view is first generated based on the DA with single-temporal samples, and then the spatial branch and the channel branch in the Hierarchical Contrastive Head are assigned to extract spatial and semantic features, respectively. The pretrained encoder possesses the ability to accurately capture geometric and categorical information of the targets, which can be directly used for downstream CD baseline fine-tuning after model clipping and alignment. The experimental results demonstrate that our SSLChange framework provides benefits to the baseline in the RSCD task, especially in the case of the data-limited situation enabling the baseline to be more stabilized in the training process and obtaining a more satisfactory performance.

In future work, we will further consider the existing limitations and explore various self-supervision-based pretraining strategies. In addition, the efficient migration of pretraining generalized frameworks to downstream tasks at the cost of smaller computational resources is another direction of our research.

REFERENCES

[1] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[2] G. Cheng et al., "Change detection methods for remote sensing in the last decade: A comprehensive review," 2023, *arXiv:2305.05813*.

[3] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sens.*, vol. 14, no. 4, p. 871, Feb. 2022.

[4] H. Jiang et al., "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 7, p. 1552, Mar. 2022.

[5] Z. Zheng et al., "EMS-CDNet: An efficient multi-scale-fusion change detection network for very high-resolution remote sensing images," *Int. J. Remote Sens.*, vol. 43, no. 14, pp. 5252–5279, Jul. 2022.

[6] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 101–119, May 2022.

[7] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.

[8] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[9] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, 2022.

[10] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[11] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.

[12] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5633520.

[13] L. Hu, Q. Liu, J. Liu, and L. Xiao, "PRBCD-Net: Predict-refining-involved bidirectional contrastive difference network for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5620717.

[14] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, Aug. 2023.

[15] L. Hu, J. Liu, and L. Xiao, "A total variation regularized bipartite network for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5239518.

[16] Y. Zhou, X. Li, K. Chen, and S.-Y. Kung, "Progressive learning for unsupervised change detection on aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601413.

[17] H. Noh, J. Ju, M. Seo, J. Park, and D.-G. Choi, "Unsupervised change detection based on image reconstruction loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1352–1361.

[18] M. Gong, J. Li, Y. Zhang, Y. Wu, and M. Zhang, "Two-path aggregation attention network with quad-patch data augmentation for few-shot scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4411616.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Jul. 2019.

[20] X. Hao, L. Liu, R. Yang, L. Yin, L. Zhang, and X. Li, "A review of data augmentation methods of remote sensing image target recognition," *Remote Sens.*, vol. 15, no. 3, p. 827, Feb. 2023.

[21] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.

[22] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[23] S. Yun, D. Han, and S. J. Oh, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6023–6032.

[24] 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. 9729–9738.

[25] 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.

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

[28] Z. Xue, X. Yu, A. Yu, B. Liu, P. Zhang, and S. Wu, "Self-supervised feature learning for multimodal remote sensing image land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533815.

[29] P. Berg, M.-T. Pham, and N. Courty, "Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives," *Remote Sens.*, vol. 14, no. 16, p. 3995, Aug. 2022.

[30] L. Jian, Z. Pu, L. Zhu, T. Yao, and X. Liang, "SS R-CNN: Self-supervised learning improving mask R-CNN for ship detection in remote sensing images," *Remote Sens.*, vol. 14, no. 17, p. 4383, Sep. 2022.

[31] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, "TOV: The original vision model for optical remote sensing image understanding via self-supervised learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4916–4930, 2023.

[32] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610426.

[33] Z. D. Calhoun, S. Lahrichi, S. Ren, J. M. Malof, and K. Bradbury, "Self-supervised encoders are better transfer learners in remote sensing applications," *Remote Sens.*, vol. 14, no. 21, p. 5500, Nov. 2022.

[34] Y. Zhao, J. Liu, J. Yang, and Z. Wu, "Remote sensing image scene classification via self-supervised learning and knowledge distillation," *Remote Sens.*, vol. 14, no. 19, p. 4813, Sep. 2022.

[35] M. Hu, C. Wu, and L. Zhang, "HyperNet: Self-supervised hyperspectral spatial–spectral feature understanding network for hyperspectral change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5543017.

[36] J. Li, M. Gong, H. Liu, Y. Zhang, M. Zhang, and Y. Wu, "Multiform ensemble self-supervised learning for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4500416.

[37] Z. Li et al., "Few-shot hyperspectral image classification with self-supervised learning," *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, 2023, Art. no. 4500416.

[38] S. Ma, L. Tong, J. Zhou, J. Yu, and C. Xiao, "Self-supervised spectral–spatial graph prototypical network for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5517917.

[39] F. Wei, Y. Gao, Z. Wu, H. Hu, and S. Lin, "Aligning pretraining for detection via object-level contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 22682–22694.

[40] M. Li et al., "AlignDet: Aligning pre-training and fine-tuning in object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2023, pp. 6866–6876.

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

[42] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," 2019, *arXiv:1909.05948*.

[43] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 20–33, May 2021.

[44] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and LiDAR data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023.

[45] J.-Y. Yang, H.-C. Li, J.-H. Yang, L. Pan, Q. Du, and A. Plaza, "Multifrequency graph convolutional network with cross-modality mutual enhancement for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5505914.

[46] 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.

[47] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.

[48] A. Chaman and I. Dokmanic, "Truly shift-invariant convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3773–3783.

[49] E. Crawford and J. Pineau, "Spatially invariant unsupervised object detection with convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 3412–3420.

[50] S. Mei, X. Chen, Y. Zhang, J. Li, and A. Plaza, "Accelerating convolutional neural network-based hyperspectral image classification by step activation quantization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.

[51] C. Zhang et al., "A domain adaptation neural network for change detection with heterogeneous optical and SAR remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 109, May 2022, Art. no. 102769.

[52] M. Chen, Y. Zhao, T. Fang, Q. Zhu, S. Yan, and F. Gao, "Geometric and non-linear radiometric distortion robust multimodal image matching via exploiting deep feature maps," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 233–240, Aug. 2020.

[53] M. Zhang, X. Zhao, W. Li, Y. Zhang, R. Tao, and Q. Du, "Cross-scene joint classification of multisource data with multilevel domain adaption network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 11514–11526, Aug. 2023.

[54] Y. Su, J. Li, A. Plaza, A. Marinoni, P. Gamba, and S. Chakravortty, "DAEN: Deep autoencoder networks for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4309–4321, Jul. 2019.

[55] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, Jun. 2016.

[56] J. Sun, J. Liu, and L. Xiao, "Intrinsic decomposition model-guided two-stream coupled autoencoder for unsupervised hyperspectral image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[57] Z. Yao, Y. Wang, H. Wu, J. Wang, and M. Long, "ModeRNN: Harnessing spatiotemporal mode collapse in unsupervised predictive learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13281–13296, Nov. 2023.
[58] Z. Zhang, J. Yang, and Y. Du, "Deep convolutional generative adversarial network with autoencoder for semisupervised SAR image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[59] G. Dong, G. Liao, H. Liu, and G. Kuang, "A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 44–68, Sep. 2018.
[60] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[61] A. Creswell et al., "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.
[62] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.
[63] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, Sep. 2021.
[64] Y. Shi, L. Han, L. Han, S. Chang, T. Hu, and D. Dancey, "A latent encoder coupled generative adversarial network (LE-GAN) for efficient hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5534819.
[65] H. Ham, T. Joon Jun, and D. Kim, "Unbalanced GANs: Pre-training the generator of generative adversarial network using variational autoencoder," 2020, *arXiv:2002.02112*.
[66] O. Manas, A. Lacoste, X. Giro-I-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9414–9423.
[67] F. Jiang, M. Gong, H. Zheng, T. Liu, M. Zhang, and J. Liu, "Self-supervised global–local contrastive learning for fine-grained change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4400613.
[68] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "CMID: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607817.
[69] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," 2022, *arXiv:2206.13188*.
[70] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618014.
[71] Y. Chen and L. Bruzzone, "Self-supervised remote sensing images change detection at pixel-level," 2021, *arXiv:2105.08501*.
[72] S. Jia, X. Zhou, S. Jiang, and R. He, "Collaborative contrastive learning for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507714.
[73] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.
[74] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[75] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.
[76] Y. Zhao, T. Celik, N. Liu, and H.-C. Li, "A comparative analysis of GAN-based methods for SAR-to-optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[77] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.
[78] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.
[79] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, May 2018.
[80] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.
[81] T. Lei et al., "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402114.

Yitao Zhao (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in photogrammetry and remote sensing from Southwest Jiaotong University, Chengdu, China, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree in information and communication engineering with the School of Information Science and Technology.

His research interests include change detection and self-supervised learning.

Turgay Celik (Senior Member, IEEE) received the Ph.D. degree from the University of Warwick, Coventry, U.K., in 2011.

He is currently a Professor with University of the Witwatersrand, Johannesburg, South Africa. His research interests include signal and image processing, computer vision, machine intelligence, robotics, data science, and remote sensing.

Dr. Celik is an Associate Editor of *IET Electronics Letters* (ELL), the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS), and Springer SIVP.

Nanqing Liu (Member, IEEE) received the B.Sc. degree in electronic information engineering from Wuhan University of Science and Technology, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree in information and communication engineering with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China.

His research interests include label-efficient object detection and optical remote-sensing image processing.

Feng Gao (Member, IEEE) received the B.Sc. degree in software engineering from Chongqing University, Chongqing, China, in 2008, and the Ph.D. degree in computer science and technology from Beihang University, Beijing, China, in 2015.

He is currently an Associate Professor with the School of Information Science and Engineering, Ocean University of China, Qingdao, China. His research interests include remote sensing image analysis, pattern recognition, and machine learning.

Heng-Chao Li (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in information and communication engineering from Southwest Jiaotong University, Chengdu, China, in 2001 and 2004, respectively, and the Ph.D. degree in information and communication engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2008.

From November 2013 to October 2014, he was a Visiting Scholar with the University of Colorado Boulder, Boulder, CO, USA, under the supervision of Prof. William J. Emery. He is currently a Full Professor with the School of Information Science and Technology, Southwest Jiaotong University. His research interests include statistical analysis of SAR images, remote sensing image processing, and pattern recognition.