Self-Supervised Learning for High-Resolution Remote Sensing Images Change Detection With Variational Information Bottleneck

Congcong Wang, Shouhang Du, Wenbin Sun, and Deqin Fan

***Abstract*—Notable achievements have been made in remote sensing images change detection with sample-driven supervised deep learning methods. However, the requirement of the number of labeled samples is impractical for many practical applications, which is a major constraint to the development of supervised deep learning methods. Self-supervised learning using unlabeled data to construct pretext tasks for model pretraining can largely alleviate the sample dilemma faced by deep learning. And the construction of pretext task is the key to the performance of downstream task. In this work, an improved contrastive self-supervised pretext task that is more suitable for the downstream change detection is proposed. Specifically, an improved Siamese network, which is a change detection-like architecture, is trained to extract multilevel fusion features from different image pairs, both globally and locally. And on this basis, the contrastive loss between feature pairs is minimized to extract more valuable feature representation for downstream change detection. In addition, to further alleviate the problem of little priori information and much image noise in the downstream few-sample change detection, we propose to use variational information bottleneck theory to provide explicit regularization constraint for the model. Compared with other methods, our method shows better performance with stronger robustness and finer detection results in both quantitative and qualitative results of two publicly available datasets.**

***Index Terms*—Change detection, contrastive learning, remote sensing, self-supervised learning, variational information bottleneck (VIB).**

I. Introduction

AGAINST the backdrop of global change, accurate and rapid sensing of surface changes is of great practical significance for ecological environmental protection, natural resource management, and for carrying out sustainable planning and governance [1]. The technology of change detection with remote sensing images has gradually become the main way for humans to discover and understand changes and has been widely used in agricultural mapping [2], land cover change detection [3], disaster detection [4], and other tasks. In addition, with the continuous development of earth observation technology, high-resolution remote sensing images with a spatial resolution of meters or even submeters are increasingly used for fine analysis due to their clearer description of ground objects [5]. However, due to the complex structure and variable scale, there are still many challenges to detect changes.

In recent years, deep learning methods have attracted increasing attention and achieved great success in the field of remote sensing images change detection. According to the number of labeled samples in the training stage, they can be divided into supervised [6], [7], unsupervised [8], [9] and semisupervised method [10], [11]. In supervised learning, sufficient labeled samples are provided for network training. And in reality, the recent development and success of deep learning method for change detection also mainly focus on supervised learning. However, in practical applications, the labeling of samples is a work that consumes manpower, material, and financial resources, especially the dense pixel-level labeling, which greatly limits the application of supervised method in practical scenarios [12]. One strategy to address this limitation is the unsupervised method, which is able to automatically learn the intrinsic distribution characteristics of the data without any labeled samples. And currently, domain adaptation [13] and image transformation or reconstruction [14], [15] are the two mainstream methods in unsupervised change detection. However, the problem of these two types of methods is that their focus is not placed on the change detection, e.g., domain adaptation methods focus on reducing interdomain differences, and image transformation or reconstruction methods pay more attention to perfect image transformation and reconstruction. In summary, the above-mentioned two methods do not consider the need for features applicable to the change detection to be sufficiently discriminative [16], [17]. On the other hand, semisupervised learning is an alternative paradigm to alleviate the constraint of labeled samples [18]. This method is able to extract latent feature information from a large number of unlabeled samples and transfer it to the downstream model, making it possible to train the network with a small number of labeled samples and solve the target task. Currently, self-supervised learning has a dominant presence in semisupervised learning, especially contrastive self-supervised learning [19], [20], [21], [22]. It has been proved to be able to extract more meaningful feature representations for downstream task by constructing positive and negative sample pairs and training the network under the constraint of contrastive loss. However, most of existing

Manuscript received 17 February 2023; revised 24 April 2023 and 30 May 2023; accepted 18 June 2023. Date of publication 21 June 2023; date of current version 7 July 2023. This work was supported in part by the General Program of National Natural Science Foundation of China under Grant 42271435 and Grant 42201512. *(Corresponding author: Shouhang Du.)*

The authors are with the College of Geoscience and Surveying Engineering, China University of Mining and Technology—Beijing, Beijing 100083, China (e-mail: bqt2000205060@student.cumtb.edu.cn; dush@cumtb.edu.cn; swb@cumtb.edu.cn; deqinfan@cumtb.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3288294

methods construct self-supervised framework with contrastive learning in the most classical way and do not consider the intrinsic relevance of contrastive learning architecture to downstream task. Therefore, it is a key to construct a contrastive learning architecture to provide a better direction of parameter optimization for downstream few-sample change detection model.

In addition, the key constraint on the performance of few-sample deep learning is that it is difficult for the model to extract sufficient prior knowledge related to label from limited labeled samples, resulting in insufficient representation ability of the extracted features. From another perspective, the geographic structure and composition in high-resolution remote sensing images are more complex and variable, and there are more pseudochanges caused by noise such as shadows, illumination, and seasonal changes. In this case, it is more difficult for the model under the supervision of few samples to focus on the real change information, which leads to a lot of information in the extracted feature representation that is irrelevant to the task of change detection. And it is not conducive to the identification of changed and unchanged regions. Therefore, it is crucial to explicitly provide certain prior information to regularize the model, suppress the introduction of noise, and make it focus on the extraction of discriminative features.

Considering the above problems, in this work, we propose an improved change detection method based on contrastive self-supervised pretraining and variational information bottleneck (VIB) theory for few-sample change detection. The motivations of the method lie in two aspects. On the one hand, the design of existing self-supervised pretext tasks does not adequately consider the characteristics of downstream tasks, which limits the potential information provided by the pretrained model. On the other hand, a much more complex scene and noise in high-resolution remote sensing images may lead to redundant and unreliable features. Based on the above motivations, a contrastive self-supervised framework oriented to the downstream change detection is designed in the proposed method. The backbone of the framework is a four-branch Siamese network, which is trained in a contrastive manner using the bitemporal images and their enhanced images. In this method, the deep feature of individual image in the classical architecture of contrastive learning is replaced by the fusion features between bitemporal images. And the contrastive learning is performed at two granularities of image global and local and two levels of shallow texture and deep semantic features, respectively. In addition, in the fine-tuning stage of downstream change detection, a regularization constraint based on the VIB theory is exploited to guide the network to focus on the changed region, reduce the interference of noise information, which further improves the final change detection results. The main contributions of this work are summarized as follows.

1) A contrastive self-supervised pretraining framework for change detection is proposed. The training is performed with fusion features of bitemporal images as the basic contrast unit to fully learn the intrinsic relationship between bitemporal images using unlabeled images, so as to provide better initialization parameters for the downstream change detection network.
2) Multilevel and multigranularity feature extraction is proposed. Emphasizing the characteristics of different levels and granularity features, contrastive learning is performed globally and locally from the perspective of deep semantic and shallow texture, and multilevel change detection is performed. In this way, the overall and detailed information can be taken into account and the detection performance can further be improved.
3) A simple but effective regularization method is proposed in this work. By adding the VIB constraint, the change detection network can focus on the extraction of real change information to weaken the interference caused by noise, thereby enhancing the effect of change detection.

The rest of this article is organized as follows. Section II briefly presents the work related to self-supervised learning and information bottleneck. Section III describes the proposed approach in detail, including the architecture of contrastive self-supervised network and change detection with the VIB. Section IV presents the results of comparison with other methods on two publicly available datasets, ablation experiments, and analysis of relevant factors. Section V provides a further discussion. Finally, the article is summarized in Section VI.

II. Related Work

In this section, we begin with an overview of the existing works on self-supervised learning and information bottleneck related to our method, which lays the groundwork for the proposal of our method in the following sections.

A. Self-Supervised Learning

The main reason why previous deep learning methods have been difficult to generalize in practical applications is that supervision of a large number of labeled samples is required, and it is impractical to collect these samples. On the contrary, there is a large amount of unlabeled data that are not effectively utilized in real scenarios. In such a case, self-supervised learning methods, which can automatically learn the intrinsic characteristics of data without any manual annotation, emerged and showed great potential. First, pseudosupervision with a large amount of unlabeled data is performed by setting pretext tasks to replace the supervision of real labels, and then, the learned latent information is transferred to the downstream task by knowledge transfer. Further, the network is trained under the supervision of a small number of labeled samples with a view to achieving comparable results to supervised learning. Due to the above characteristics, self-supervised learning methods have attracted more and more attention in image classification, segmentation, and change detection [23], [24], [25], [26], [27].

The key to self-supervised learning method is the construction of pretext task, which can be roughly divided into contrastive method and context-based method. Among them, the noncontrastive pretext tasks such as grayscale image recoloring [28], image rotation angle prediction [29], relative position of image blocks prediction [30], and image inpainting [31], which take a single image as input, have been shown to be effective for some specific downstream tasks. To obtain more sufficient

latent features, there are also methods that integrate the above different tasks for self-supervised learning and achieve good results [32], [33], [34], [35]. On the other hand, contrastive self-supervised learning constructed in a contrastive manner can extract discriminative features by making positive sample pairs close to each other and negative sample pairs away from each other. The mode of construction and the optimization direction of this pretext task are far more relevant to the change detection than the above-mentioned noncontrastive pretext tasks. More importantly, it has been shown that features extracted by contrastive learning are more favorable to downstream task optimization and largely narrow the gap with supervised methods [36], [37], [38]. The basic context of this method is, first, the generation of sample pairs. Specifically, the samples generated by the image transformation without changing the semantic are taken as positive samples, and the other samples are regarded as negative samples. Then the distance between sample pairs is measured and constrained, and the network is further trained. On this basis, there are methods that keep the architecture of contrastive learning unchanged and learn more efficient feature representation by changing the way of constructing positive and negative sample pairs. For example, Dwibedi et al. [39] took the closest samples in the embedding space as positive samples to learn key knowledge from more relevant objects. Starting from the basic properties of hyperspectral images, Lee and Kwon [25] considered images blocks in adjacent regions as positive samples with similar spectral properties and images blocks in different regions as negative samples. In addition, there are some studies that construct positive and negative sample pairs through similarity measurement [24], [26]. Recently, contrastive learning combined global and local contrasts and proved to be more suitable for the downstream task of pixel-level change detection. Jiang et al. [40] followed the general contrastive learning framework to perform global and local contrastive pretraining with a single image as the contrast unit. And the change detection backbone network is proposed as the feature extractor in the pretraining network to adapt to the downstream change detection. Inspired by these methods, and considering advantages of contrastive learning and its fit with the change detection, it is used as the basic methodological framework for pretraining. We extend the image-level contrastive method and improve the network architecture to extract discriminative features from different feature levels and granularity for downstream task.

B. Information Bottleneck Theory

Information bottleneck theory [41] is extended from rate distortion theory in the field of data compression to find an optimal representation between source distortion and rate reduction. In 2016, Alemi et al. [42] proposed variational inference to solve the problem of difficult mutual information calculation in the information bottleneck, i.e., VIB, and further linked it with deep learning. Currently, the applications of information bottleneck and VIB in the field of deep learning mainly focus on feature representation and interpretable deep learning. In [43], [44], and [45], the attention mechanism based on VIB theory is proposed, and the theory has been shown to be effective in extracting key features of interest and reducing the interference of irrelevant information. In reinforcement learning and natural language processing [46], [47], [48], the addition of VIB constraint also significantly enhances the feature representation in low resource. In [49], VIB theory is applied in multiview representation learning. By maximizing the mutual information between the desired representation and the shared representation, maximizing the mutual information between the desired representation and view-specific representation, and minimizing the mutual information between the desired representation and the original image, the shared and specific information between different views is decoupled and irrelevant information is filtered out to explore the optimal complete representation. In addition, the graph neural network based on VIB is also proposed to explore the key graph structure and improve the quality of the final graph representation in the application of unstructured data [50]. In terms of interpretable deep learning, Schulz et al. [51] added a bottleneck layer to the network to reveal the importance of different regions of an image to the network prediction by adding noise, limiting the flow of information, and observing the prediction results. In general, VIB theory can limit the flow of irrelevant information to a certain extent and make the network focus on the key features of interest. In the change detection of high-resolution remote sensing images focused on in this work, the noise, such as illumination, shadows, and seasonal changes, will bring adverse effects. And reducing the effect of noise is expected to improve the accuracy of change detection. Therefore, we consider incorporating this theory into the network framework to alleviate the negative impact of irrelevant information. Moreover, as far as we know, this theory has not been discussed in the field of remote sensing images change detection so far.

III. Method

In this section, we detail the proposed change detection method for high-resolution remote sensing images with a small number of labeled samples. First, the overall framework of the method is introduced, and then its two important components, contrastive self-supervised pretraining and change detection with VIB, are described in detail.

A. Overall Framework

Our approach is designed for applications where the dataset contains a small number of labeled samples and a large number of unlabeled samples. The whole training framework is shown in Fig. 1 and consists of two stages: a contrastive self-supervised pretraining stage and a few-sample change detection fine-tuning stage. The purpose of the pretraining is to fully explore the latent knowledge of a large number of unlabeled samples to provide a good direction for parameter optimization in the fine-tuning stage. First, both labeled samples and unlabeled samples are taken as input. Then, the features of the bitemporal images and their corresponding enhanced samples are extracted separately by a four-branch Siamese network, and the network parameters are optimized by constraining the extracted multilevel and multigranularity features of bitemporal images using contrastive

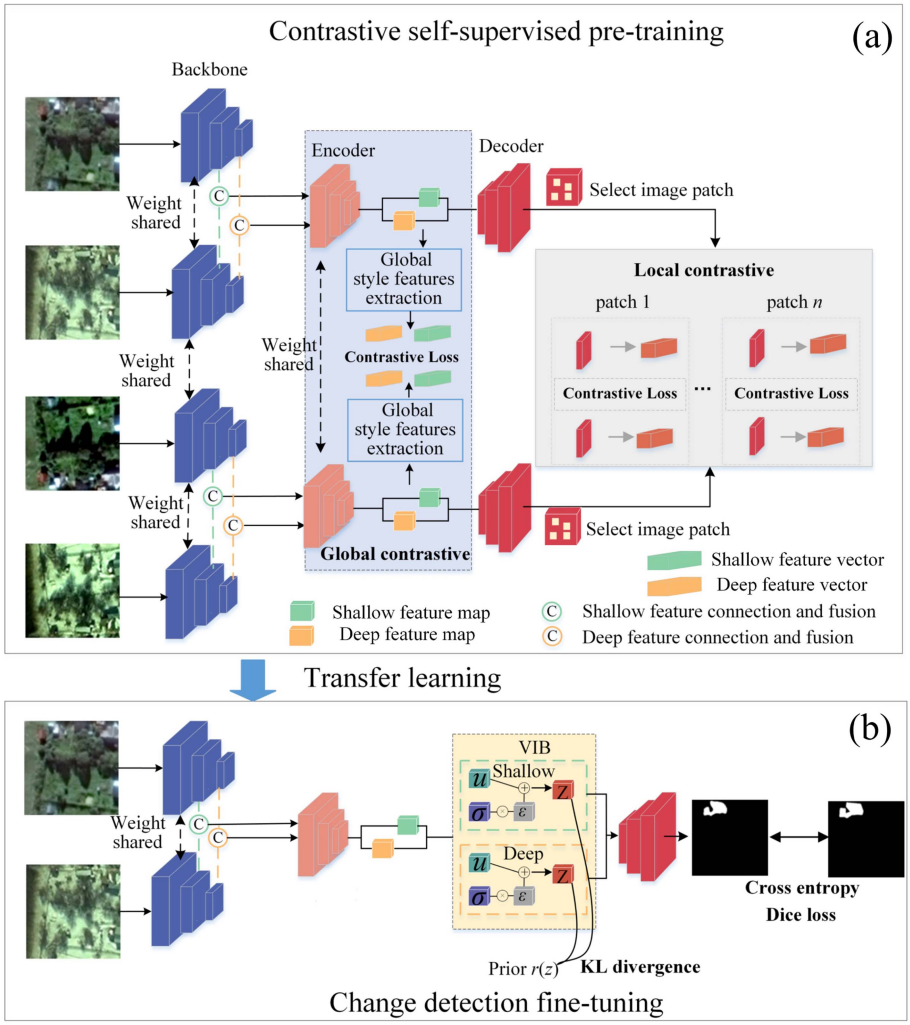

Fig. 1. Illustration of the proposed change detection approach. (a) Contrastive self-supervised pretraining: the four-branch Siamese network takes bitemporal images before and after image enhancement as input and performs multilevel and multigranularity contrastive pretraining. (b) Change detection fine-tuning: the pretrained model parameters are transferred to the downstream task using transfer learning, and the change detection model is fine-tuned using few samples under the constraint of VIB.

loss. In the fine-tuning stage, the pretrained network parameters are first transferred to the change detection network through knowledge transfer, and then the network is fine-tuned for multilevel change detection under the supervision of a small number of labeled samples. In addition, to further increase the prior knowledge of network and reduce the interference of irrelevant noise, the VIB theory is introduced to regularize the model of change detection. Finally, in the test stage, the test images are fed into the trained change detection network to generate change maps directly in an end-to-end manner.

B. Contrastive Self-Supervised Pretraining

In order to learn latent knowledge suitable for bitemporal images change detection from unlabeled images, the proposed contrastive self-supervised pretraining network integrates a combination of global and local, shallow and deep feature-level contrastive learning. A four-branch Siamese network with parameter sharing is used, which maps bitemporal images and their enhanced images to the same feature space and then compares the fusion features of bitemporal images at different feature levels and granularity to extract discriminative features. Specifically, the whole pretraining stage consists of two steps.

First, the backbone network ResNet50 is used to extract the features of the four images respectively and output the corresponding abstract features of two different levels, and the features are divided into shallow and deep features, and further a two-branch encoder is constructed for late fusion and change feature extraction for shallow and deep features of different time phases, respectively. Finally, the network is trained using the contrastive loss function to constrain the feature distance between positive and negative sample pairs. The global and local sample pairs are acquired, as shown in Fig. 2. Specifically, the bitemporal image pair is used as the contrast unit for pretraining. And global contrast emphasizes on distinguishing image samples from other different image samples so that the enhanced image corresponding to the original image is used as the positive sample and the other samples are the negative samples. Local

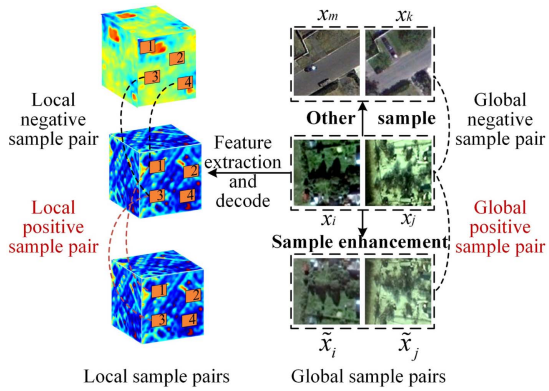

Fig. 2. Construction of positive and negative sample pairs. With the bitemporal images as the contrast unit, global positive and negative sample pairs are images before and after image enhancement, and local positive and negative sample pairs are image blocks.

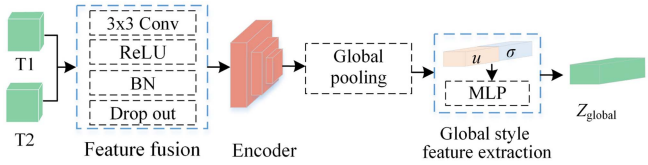

Fig. 3. Structure of feature fusion and deeper feature extraction. The bitemporal feature maps are the input, and the fused global feature vector is the output.

contrast focuses on the local information of the image and differs from the former in that it takes the image block as the contrast unit. The image blocks in the corresponding positions of the original image and enhanced image are positive samples, and the other blocks in different images and different regions are negative samples. The details of the feature extraction network and the contrastive loss function are as follows.

1) Multilevel and Multigranularity Discriminative Feature Extraction: The recent development and success of bitemporal images change detection methods mainly focuses on the Siamese network architecture with late fusion [6]. The reason for this is that, in the early stage, the two-branch network can separately extract features from the bitemporal images and retain the original features of each image. And the late fusion can fuse and discriminate the extracted features on higher dimension and further improve the situation that the low-dimensional feature space is not separable. Inspired by this, four-branch Siamese network is constructed to improve the problem that existing two-branch self-supervised network may overlook the intrinsic relationship between bitemporal images. Specifically, the four-branch network is used to extract features from the bitemporal images and their enhanced images separately, and then the extracted bitemporal image features are fused and deeper feature extraction are carried out. Moreover, on the one hand, it is considered that there are not only global differences, such as weather and illumination, but also local differences at pixel level in the bitemporal images change detection. Therefore, inspired by Li et al. [52], we perform feature extraction from both global and local. On the other hand, in deep networks, shallow features are beneficial to detect changes in low-level image texture and locate the spatial location of the changed regions. Deep features are beneficial to detect changes in high-level semantic level and identify whether changes are present or not. Considering the above facts, we propose multilevel and multigranularity feature extraction architecture that combines global and local, shallow and deep features. The computational details are as follows.

As shown in Fig. 1, in the global contrastive module, feature extraction is first performed on single image using the backbone network ResNet50, which includes five stages of feature extraction. And the feature maps output in the second stage are regarded as shallow feature, focusing on texture information. Correspondingly, the final output feature maps of the fifth stage of the network are regarded as deep feature for feature analysis at the semantic level. After that, in order to establish the internal relationship between the bitemporal images and extract discriminative features that can reveal changes, feature fusion and deeper feature extraction are performed on the shallow and deep features of bitemporal images, respectively. The two processes are the same, as shown in Fig. 3. The difference lies in that the inputs are bitemporal shallow features and deep features, respectively. Specifically, the shallow and deep features of bitemporal images are connected along the channel, and then fed into the feature fusion module and encoder for late fusion and feature extraction, respectively. The feature fusion module contains a 3×3 convolution, a batch normalization layer, a nonlinear activation layer ReLU, and a dropout. And the encoder is DeepLabV3+, whose role is to encode the fusion features output from the feature fusion module at a deeper level. After that, the global pooling is performed on the features output from the encoder, and then the mean and variance of the pooled features are calculated, and the two are connected as the input of the single hidden layer MLP. Finally, the nonlinear global feature vector Z_{global} at different levels is obtained.

In addition, the extraction of local features is shown in the local contrastive module in Fig. 1. First, the shallow and deep features output from DeepLabV3+ are fed into the corresponding decoder network to obtain the feature map with the same size as the original image. Then, as shown in the local sample pairs' selection strategy in Fig. 2, local features of multiple regions are randomly selected from the decoded feature map. Finally, in a similar way to the global style feature vector extraction, the local feature vector Z_{local} is obtained by performing global pooling and MLP projection on the local features obtained in the previous step.

2) Contrastive Loss Function: The contrastive loss function is very effective in the processing of paired data and is an essential part of the contrastive learning. In this work, we use the classical Info NCE [53] as the contrastive loss function, which is represented as follows:

$$L_{\mathrm{C}} = \frac{1}{2N} \sum_{i=1}^{N} \left(l\left(x_i, \tilde{x}_i\right) + l\left(\tilde{x}_i, x_i\right) \right) \tag{1}$$

$$l\left(x_i, \tilde{x}_i\right) = -\log \frac{\exp\left(\operatorname{sim}\left(z_i, \tilde{z}_i\right)/\tau\right)}{\sum_{k=1, k\neq i}^{2N} \exp\left(\operatorname{sim}\left(z_i, z_k\right)/\tau\right)} \tag{2}$$

where N represents the total number of original samples, which is expanded to $2N$ after enhancement. And the sample and its corresponding enhanced sample form one positive sample pair, and the other samples are negative samples, the number is $2(N\text{-}1)$. x_i and $\tilde{x}_i$ represent the images before and after sample enhancement, and z is the extracted feature representation, corresponding to the global and local feature vectors in multilevel and multigranularity discriminative feature extraction. τ is a hyperparameter, which is used to adjust the discrimination of the model for negative samples. sim is used to measure the distance between samples, and in this work, the Euclidean distance and cosine distance are, respectively, used for training, and the results show that the model trained based on the cosine distance can provide better initialization parameters for downstream change detection, which makes the corresponding change detection better. Therefore, the cosine distance is used in all subsequent experiments. As can be seen from (2), the numerator in the contrastive loss function is used to measure the distance between positive sample pairs, while the denominator calculates the distance between all negative sample pairs. Minimizing the loss function can promote the similarity between positive sample pairs to be as high as possible and the similarity between negative sample pairs to be as low as possible so that the network under this constraint can extract discriminative information.

C. Change Detection With VIB

After the contrastive self-supervised pretraining, the pretrained network learns some prior knowledge related to downstream change detection. To make full use of this advantage and provide a better search direction for the subsequent change detection network, we construct a fine-tuning network similar to the pretraining network, and the relationship between the two is shown in Fig. 1. Specifically, we use the backbone and bitemporal feature fusion encoder consistent with the pretraining network and take the parameters of the pretrained network as the initialization parameters of the fine-tuning network through knowledge transfer. And then, the multilevel change detection network is trained under the supervision of a small number of labeled samples.

However, although the change detection network already contains part of the prior information introduced by the contrastive self-supervised learning, there are still some problems. That is, the network is still limited by a small number of labeled samples, resulting in a lot of information in the extracted features irrelevant to the label and poor detection effect. Therefore, we propose to introduce the theory of VIB to provide a more explicit optimization direction for the model.

The goal of VIB is to learn a feature representation Z that maximizes the preservation of information related to the label Y while compressing the information of input X and reducing irrelevant information. This is consistent with two properties that the change detection model should satisfy: 1) the predicted change map should be as similar as possible to the ground truth; 2) the focus should be on the real change characteristics and the interference of irrelevant information should be as little as possible. For this, the optimization objective can be expressed as minimizing the following equation:

$$L = \beta I(X, Z) - I(Z, Y) \tag{3}$$

where $I(,)$ represents the mutual information, X corresponds to the input image data, Y is the corresponding ground truth, and Z is the optimal feature representation to be extracted. $I(X, Z)$ is used to measure the mutual information between the input source X and the compressed feature representation Z. $I(Z, Y)$ represents the correlation degree between the feature representation Z and the target Y. Minimizing (3), i.e., maximally compressing X by discarding the information irrelevant to Y, minimizing $I(X, Z)$. Simultaneously, maximizing $I(Z, Y)$ to capture as much information relevant to Y as possible to obtain the optimal feature representation of the target Y. β is used to tradeoff between information compression and label prediction.

The difficulty is that it is difficult to calculate the mutual information of high-dimensional features in (3). However, in this work, we actually do not care about the corresponding exact value but just need to find the optimal solution by minimizing (3). Therefore, we use the variational estimation proposed in [42] to solve the upper bound of (3). Assuming that $q(y|z)$ and $r(z)$ are variational approximation of $p(y|z)$ and $p(z)$, respectively, we can obtain

$$L \leq \beta \mathrm{E}_{p_{x,z}}\left[\log \frac{p(z|x)}{r(z)}\right] - \mathrm{E}_{p_{x,y,z}}[\log q(y|z)]. \tag{4}$$

Then, the reparameterization trick [54] is used to express z as $z = \mu(x) + \sum(x) \times \varepsilon$, where $\varepsilon \in N(0, I)$. The following can be obtained:

$$L \leq \beta \mathrm{KL}\left[p(z|x) \,|r(z)\right] - \mathrm{E}_{p_{x,y,\varepsilon}}[\log q(y|f(x, \varepsilon))] \tag{5}$$

where f in the second term on the right side of the inequality represents the network mapping used to identify changes, while the entire second term is the cross entropy and represents the prediction error. In addition, class imbalance is a common problem in practical change detection. Dice loss, as a region-related loss function, has been shown to be effective in alleviating the class imbalance problem and can produce complementary effects [55], [56] when combined with loss functions, such as cross entropy. Therefore, we combine dice loss and use both as the evaluation function of the change detection error due to the problem of class imbalance. In addition, in the calculation of KL divergence, we assume that $p(z|x)$ and $r(z)$ obey the Gaussian distribution. Now, the final representation of the optimization objective is as follows.

$$L \leq \beta\left(-\sum_{k=1}^{K}\log(\sigma_k) + \frac{1}{2}\sum_{k=1}^{K}\left(u_k^2 + \sigma_k^2\right) - \frac{K}{2}\right) - \left(L_{\text{bce}} - \log(L_{\text{dice}})\right) \tag{6}$$

where K is the dimension of the extracted feature, and σ_k and u_k are the standard deviation and mean of the feature, respectively. L_{bce} and L_{dice} represent the cross entropy and dice loss, respectively, where log is used to adjust both to the same scale. Accordingly, the whole fine-tuning network is guided by (6) and is trained under the supervision of a small number of

labeled samples. Finally, excellent change detection model can be obtained in this way.

IV. Experiments and Analysis

In this section, we first provide a brief description of the image data, relevant experimental settings, the comparison methods, and the evaluation metrics. Then we perform comparison experiments and ablation experiments on different datasets. In addition, the results of the above-mentioned experiments and the relevant factors in the experiments are further analyzed and discussed to fully demonstrate the effectiveness of the proposed method.

A. Experimental Datasets

To verify the feasibility of the proposed method on different datasets, we conduct experiments on two publicly available datasets. The detailed description of the datasets is as follows.

1) Lebedev Dataset [57]: The high-resolution dataset is obtained from Google Earth. Specifically, it contains three visible bands of red, green, and blue, with a maximum spatial resolution of 3 cm/px and a minimum of 100 cm/px. After processing by data enhancement methods such as rotation, translation, and color transformation, there are 12 998 pairs of 256 $\times$ 256 labeled samples in the dataset, and the labels are manually annotated by professionals. Among them, the changes to be detected include the appearance and disappearance of small cars, large-scale land cover change, and other changed objects at different scales. It is worth noting that the dataset is collected in different seasons and contains obvious seasonal differences.

2) SenseTime Dataset [58]: This high-resolution dataset is collected for the Sensetime AI Remote Sensing Interpretation Competition and contains 2968 pairs of 512 $\times$ 512 labeled samples, corresponding to a spatial resolution of 0.5 $\times$ 3 m. Similarly, the images in this dataset contain only red, green, and blue bands. In addition, the type of change focused on in this dataset is the conversion between six different land use types.

B. Experimental Settings

1) Implementation Details: All experiments are implemented on a PC platform with 12th Gen Intel(R) Core(TM) i7-12700F 2.10 GHz CPU and NVIDIA GeForce RTX 3090 graphics card, using pytorch as the underlying deep learning framework. In both stages of network training for the contractive self-supervised and change detection, the Adam optimizer is used, and the initial learning rate is set to 0.001, and the learning rate is adjusted with Cosine annealing during the training process. Besides, it is verified that the downstream change detection model has optimal performance when τ of the contrastive loss function is set to 0.5 in contrastive self-supervised pretraining, which is consistent with that in [22]. Therefore, τ is set to 0.5 in all subsequent experiments. Sample enhancement in the training process includes random cropping, resizing, flipping, rotation, color transformation, and Gaussian blur. And the two stages are trained 400 epochs and 300 epochs, respectively. The batch size of the Lebedev and SenseTime dataset in the pretraining stage is set to 32 and 6, respectively, and that of the fine-tuning stage is set to 8. In addition, to fully verify the performance of the proposed method with a small number of labeled samples, all data in both datasets are used for self-supervised pretraining, while only a small number of labeled samples are used to train the change detection fine-tuning network. Specifically, in the experiments to analyze the influence of the number of labeled samples on the change detection performance, 1%, 10%, and 20% of the labeled samples are selected to train the change detection fine-tuning network. Other than that, 1% of the labeled samples are selected as training data for all experiments. And the test set of all experiments contains 20% of the labeled samples randomly selected from both datasets. For fairness, consistent training and testing samples are used for all methods.

2) Comparison Methods: In order to show the superiority of the proposed method, seven state-of-the-art change detection methods are selected to compare with our method. Among them, the backbone and encoder of SimCLR, MoCo v2 and BYOL are consistent with our method, which are ResNet50 and DeepLabV3+, respectively. In addition, the first four of the comparison methods adopt the same network architecture in the downstream change detection task. The details of these methods are described as follows.

a) *Random initialization:* Without pretraining, the parameters of the fine-tuning network are randomly initialized using the kaiming initialization method [59].

b) *SimCLR pretraining [22]:* This method is a contrastive self-supervised method at the instance level, which trains the network with the goal of reducing the distance between positive sample pairs and increasing the distance between negative sample pairs. And feature extraction and projection of images processed by data enhancement are carried out by encoder with parameter sharing and single layer MLP, and the network is trained under the constraint of contrastive loss. Of which the data enhanced samples and the corresponding original samples form a positive sample pair, while other different samples are negative samples.

c) *MoCo v2 pretraining [60]:* Both this method and the SimCLR method mentioned above are essentially contrastive self-supervised methods. The difference is that the MoCo v2 constructs a larger queue of negative samples in a dynamic update manner to make the model learn more negative sample information. It also proposes to update the encoder using a type of momentum update.

d) *BYOL pretraining [21]:* This is an implicit contrastive learning method without negative samples, which constructs regression constraint with the target network by adding an MLP to predict the projection of online network, so as to achieve consistent prediction between positive sample pairs.

e) *FC-EF [61]:* Fully convolutional early fusion is a classical supervised change detection method. The model takes bitemporal images, which are concatenated along the channel as input and is trained with U-Net as the basic architecture and cross entropy as the loss function.

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE LEBEDEV DATASET

Type	Method	1% of the labeled sample			
		Precision (%)	Recall (%)	F1 (%)	IoU (%)
supervised method	FC-EF	37.13	31.32	32.27	19.96
	ChangeFormer	59.50	50.17	54.44	37.40
Self-supervised method	Random initialization	63.30	43.10	47.71	34.21
	SimCLR	68.04	**61.02**	62.57	47.38
	MoCo v2	62.78	58.35	58.03	44.19
	BYOL	55.71	50.72	51.25	36.28
	Multiview	11.33	51.82	18.60	10.25
	Ours	**75.23**	60.44	**65.79**	**50.61**

The bolded entities in tables represent the optimal result under the corresponding evaluation indicators.

f) *Multiview [17]:* This method is a remote sensing image change detection method based on contrastive self-supervised pretraining. In the contrastive self-supervised pretraining stage, the bitemporal images of the same region are regarded as positive sample pairs and others are negative sample pairs. And the discriminative features are extracted in a global contrastive manner. On this basis, change detection is achieved by measuring the feature distance of bitemporal images and performing threshold analysis.

g) *ChangeFormer [62]:* This method is a supervised change detection method. The model combines transformer and MLP to capture multiscale long-range information to achieve high-precision remote sensing images change detection.

3) Evaluation Metrics: To evaluate the performance of the proposed model, we use four metrics commonly used to evaluate the performance of change detection, namely Precision, Recall, F1-score, and intersection over union (IoU), which are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1} = 2 \times \frac{\text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}} \tag{9}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{10}$$

where TP indicates that the prediction result is consistent with the label, which are both changed pixels. FP indicates that the unchanged pixels are incorrectly predicted as changed pixels. Correspondingly, FN indicates that the changed pixels are incorrectly predicted as unchanged pixels.

C. Experimental Results

In this section, we present the experimental results of our method and the comparison methods on two datasets, and illustrate the superiority of our method from both quantitative and qualitative perspectives.

1) Comparison on Lebedev Dataset: The quantitative evaluation result on Lebedev dataset is given in Table I. It can be seen that the overall performance of our method outperforms the others. Among them, the method of randomly initializing network parameters, which has not been pretrained with pretext task, has a relatively poor performance, and the corresponding precision, recall, F1, and IoU are 63.30%, 43.10%, 47.71%, and 34.21%, respectively. The results of the Multiview are not satisfactory, mainly due to the fact that the downstream change detection is realized by feature differencing and threshold analysis rather than network training. It is difficult to accurately distinguish the changed regions from the unchanged regions by using the features extracted with the pretraining parameters. Compared with the above two methods, other methods based on contrastive self-supervised pretraining all achieve greater performance gains. This also implies that the pretext task pretrained in a contrastive manner can capture some additional discriminative features from the unlabeled data that are beneficial for downstream change detection, and transferring the pretrained parameters to the downstream network can significantly improve downstream task performance. And it can be seen from the results of the methods based on contrastive self-supervised pretraining, including the proposed approach, that there are great differences in the impact of different contrastive self-supervised network architectures on the performance of downstream task. Among them, the comprehensive performance of our method is the best, with precision, recall, F1, and IoU of 75.23%, 60.44%, 65.79%, and 50.61%, respectively. It is worth noting that the precision, F1, and IoU indexes are optimal in all methods. Moreover, although the highest recall of SimCLR is 61.02%, which is 0.58% higher than our method, it is much worse than

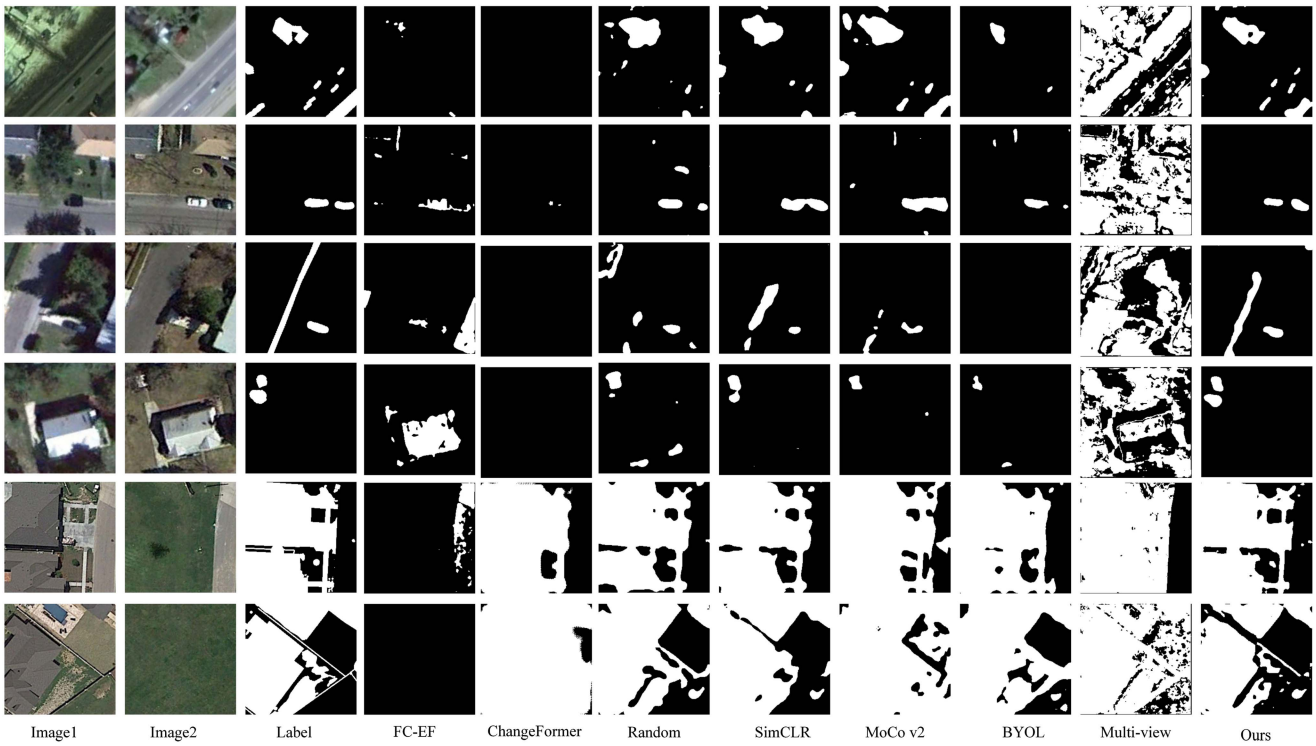

Fig. 4. Qualitative results of different methods on the Lebedev dataset. Black indicates unchanged regions and white indicates changed regions.

our method in terms of precision. It is also worth noting that compared with the FC-EF supervised method, which takes 130 labeled samples as the complete training set, the precision, recall, F1, and IoU of our method are improved by 38.10%, 29.12%, 33.52%, and 30.65%, respectively. This shows that it is difficult to extract effective features for the supervised method with simple structure when there are only a few labeled samples, which seriously affects the accuracy of change detection. Compared with the ChangeFormer, our method improves 11.35% and 13.21% in F1 and IoU, respectively; this also further validates the effectiveness and necessity of introducing self-supervised pretraining before few-sample change detection.

To better visualize the superiority of the proposed method, we show the visualization results of all methods on the test set in Fig. 4, where black and white indicate the unchanged and changed regions, respectively. As can be seen from the figure, there are many missed and false detections in the change map obtained by the FC-EF and ChangeFormer, and the overall detection effect is not satisfactory. There are many false detections in the change maps of the Multiview, which identify shadow, vegetation changes, and other pseudochange as changed regions. This indicates that in high-resolution remote sensing images, it is difficult to identify changed regions of interest by relying only on the feature representation extracted by the contrastive pretraining. Except for that, the change maps obtained by other methods are more consistent with labels, but there are also some missed and false detections. Among them, the random initialization method has relatively more false alarms, such as mistakenly treating the shadow as changes in the detection results of the second and third rows in Fig. 4. And compared with the BYOL, SimCLR and MoCo v2 have higher accuracy in identifying the changed regions and can identify most changed regions with relatively fewer missed detections. However, there is a lot of noise in the prediction results of the above three methods, which incorrectly classify many unchanged pixels as changed pixels. As in the last two rows, the above methods are not ideal for the distinction between changed buildings and unchanged roads and grass, resulting in the false detection of many unchanged pixels. The proposed method in this work can suppress the influence of noise, reduce the false and missed detections to a certain extent, and get the prediction result closer to the label. Specifically, as shown in the first row, our method can effectively identify the unchanged gaps between multiple buildings. And in the last two rows, the boundary between the changed and unchanged pixels is also clearer. All of the above-mentioned results demonstrate the effectiveness of the proposed contrastive self-supervised pretext task and the VIB constraint.

2) Comparison on SenseTime Dataset: The quantization result on the SenseTime dataset is given in Table II. We can observe that the performance on the SenseTime dataset is basically the same as that on the Lebedev dataset. Under the supervised training with 1% of all samples, i.e., 30 labeled samples, the FC-EF performs the worst. Although it has the highest precision, it also corresponds to the lowest recall. The reason for this is that under the supervision of a small number of labeled samples, the model focuses on regions with significant color or shape changes, which make the model to have relatively high precision in identifying such changes. However, in this case, the model cannot extract enough semantic features, which makes it difficult to identify changes of land use types, resulting in low recall. The ChangeFormer is a change detection model based on transformer. Compared with convolutional neural network, it has

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE SENSETIME DATASET

Type	Method	1% of the labeled sample			
		Precision (%)	Recall (%)	F1 (%)	IoU (%)
Supervised method	FC-EF	**56.61**	15.58	23.06	14.21
	ChangeFormer	20.35	47.08	28.42	16.56
Self-supervised method	Random initialization	40.82	47.19	42.70	28.17
	SimCLR	42.67	63.39	49.76	34.18
	MoCo v2	45.97	52.39	47.97	32.51
	BYOL	43.84	59.48	49.25	33.68
	Multiview	26.61	**72.30**	38.90	24.15
	Ours	49.33	57.57	**52.21**	**36.58**

The bolded entities in tables represent the optimal result under the corresponding evaluation indicators.

a larger number of parameters and usually requires a large number of data training to take advantage of it. With only 30 training samples, the correct induction bias cannot be obtained, resulting in poor performance. The results of the above-mentioned two methods indicate that the fully supervised method is unable to obtain sufficient prior knowledge from a small number of labeled samples, resulting in the model being unable to form a description of the entire data and lacking the power of discrimination. In general agreement with the results on Lebedev dataset, the Multiview has a higher recall but poorer precision. Other methods based on contrastive pretraining have the best overall performance, which confirms the claim that pretraining methods can bring improvement for downstream tasks. It can be seen that the F1 of the proposed method is 52.21%, which is an average improvement of 26.47% and 5.74% over the supervised method and other methods based on contrastive self-supervised pretraining, respectively. And the IoU also increased by 21.20% and 5.45%, respectively. It shows that the proposed contrastive self-supervised pretraining architecture is more suitable for the scenario where the downstream task is change detection and can effectively transfer the prior knowledge from unlabeled samples to the downstream network to improve the detection accuracy. Moreover, the proposed VIB regularization method provides explicit selectivity constraint for the model, making the features learned under this additional guidance more focused on the change information provided by the labels, further improving the overall change detection performance. In addition, compared with Table I, the detection performance on the SenseTime dataset is worse than that on the Lebedev dataset. This mainly has two aspects of reasons. On the one hand, the training sample size of SenseTime and Lebedev differs greatly, which largely affects the detection performance of the model on the two datasets. On the other hand, compared with the Lebedev dataset, the training objective on the SenseTime dataset is to identify the changes of different land use types, which is more diverse and complex on the images. Therefore, there are some differences in the results on the two datasets, and the overall performance on the SenseTime dataset is lower than that on the Lebedev dataset.

Fig. 5 shows the performance of the different methods on the SenseTime dataset. In general, the result of FC-EF method is unsatisfactory. And the change maps predicted by the ChangeFormer cannot reflect the changed and unchanged regions basically. This demonstrates that the training results of the supervised method depend on the number of labeled samples, and the detection performance cannot meet the requirements of practical applications when there are few available labeled samples. And the method based on transformer has higher requirements on sample size, which makes it difficult to apply to few-sample change detection. Since the Multiview uses feature difference and threshold analysis to identify changed regions, it is difficult to distinguish pseudochange from semantic change of interest, and there are many false detections. In addition to the above-mentioned methods, the change maps predicted by other methods have a good overall performance and can basically locate the changed regions. However, except for the method in this work, other methods have relatively more false detections, which is consistent with the low precision and high recall of these methods in Table II. Specifically, they cannot locate the real changed regions under the influence of scale, view, illumination, and other factors, and this problem has been improved to some extent in the proposed method. For example, in the first and fourth rows, compared with other methods, our method can detect the real changed land types and buildings under the influence of noise. Overall, the proposed method is more robust under the influence of pseudochanges and can effectively balance the precision and recall obtaining the best change map.

D. Ablation Experiments

1) Ablation Experiment of the Pretraining Model: In order to take into account the global and local, deep and shallow

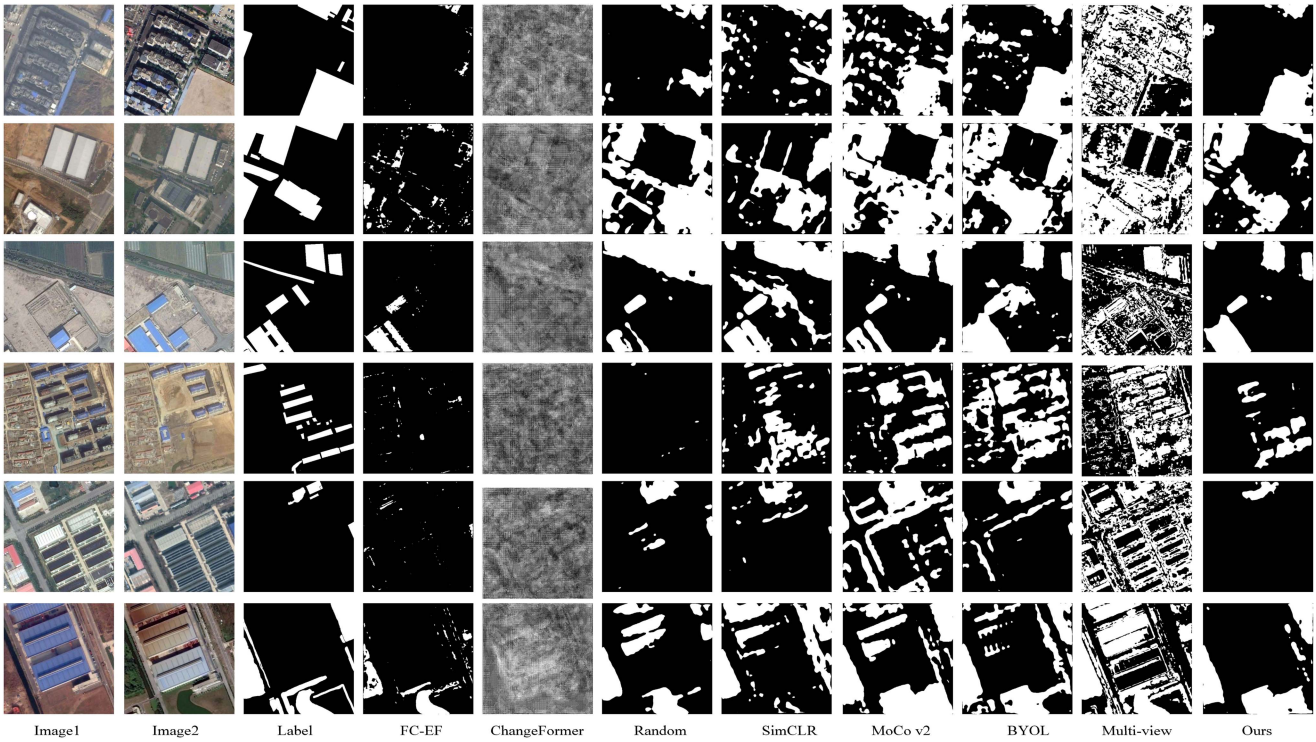

Fig. 5. Qualitative results of different methods on the SenseTime dataset. Black indicates unchanged regions and white indicates changed regions.

TABLE III
ABLATION EXPERIMENT OF THE PRETRAINING MODEL ON TWO DATASETS

Method	Lebedev Dataset				SenseTime Dataset			
	Precision (%)	Recall (%)	F1 (%)	IoU (%)	Precision (%)	Recall (%)	F1 (%)	IoU (%)
N_shallow	69.16	55.20	60.03	44.15	46.98	48.76	46.67	31.37
N_Global	68.74	**62.97**	64.35	48.54	50.20	55.89	51.58	36.01
N_Local	71.01	57.56	61.74	46.27	**50.48**	48.83	48.08	33.15
Ours	**75.23**	60.44	**65.79**	**50.61**	49.33	**57.57**	**52.21**	**36.58**

The bolded entities in tables represent the optimal result under the corresponding evaluation indicators.

features, the proposed contrastive pretraining network integrates multilevel and multigranularity contrastive learning to provide sufficient priori knowledge for downstream change detection. To verify its effectiveness, a series of ablation experiments are conducted, and the results are reported in Table III. And "*N*_shallow" indicates that only the deep features of global and local contrasts are retained in the pretraining network. Correspondingly, "*N*_Global" and "*N*_Local" indicate that only local and global contrasts are retained, respectively. It should be noted that since the backbone and feature fusion encoder of the change detection fine-tuning network is consistent with those of the pretraining network, the change detection model of "*N*_shallow" only contains change detection of deep level, while the other two models are multilevel change detection with a combination of deep and shallow levels.

As can be seen from Table III, the combining contrastive pretraining of different levels and granularity can effectively improve the model performance on both datasets, and the proposed complete model achieves optimal result. In addition, the model without shallow contrastive module has relatively poor performance on both datasets, which indicates the importance of shallow texture information in locating changed regions. On the other hand, the results from the models without global and local contrastive module reveal that the importance of local contrastive information for change detection is more prominent. This may be due to the fact that on both datasets, the changed regions are mostly represented as local change in the image, and the global feature has a relatively weaker impact on the recognition of changes. These quantitative results demonstrate not only the effectiveness of the different modules but also the gain effect of their combination.

TABLE IV
ABLATION EXPERIMENT OF VIB REGULARIZATION TERM ON TWO DATASETS

Method	Lebedev Dataset				SenseTime Dataset			
	Precision (%)	Recall (%)	F1 (%)	IoU (%)	Precision (%)	Recall (%)	F1 (%)	IoU (%)
N_VIB	74.07	59.07	64.49	48.96	48.87	53.75	50.11	34.57
Ours	**75.23**	**60.44**	**65.79**	**50.61**	**49.33**	**57.57**	**52.21**	**36.58**

The bolded entities in tables represent the optimal result under the corresponding evaluation indicators.

2) Ablation Experiment of the VIB: In this section, to further illustrate the role of the regularization term, the VIB, the ablation experiments on two datasets are conducted. Specifically, only the change detection loss combined with cross entropy and dice loss is used to train the network. And the rest settings are consistent with the proposed method. In addition, for simplicity, the change detection model without the constraint of VIB is denoted by "*N*_VIB" in the subsequent experiments.

It can be observed from the results in Table IV that the addition of VIB constraint can improve the overall model performance. More specifically, with the addition of VIB, the precision, recall, F1, and IoU of Lebedev dataset are improved by 1.16%, 1.37%, 1.30%, and 1.65%, respectively. Correspondingly, the improvements on SenseTime dataset are 0.46%, 3.82%, 2.10%, and 2.01%, respectively. These quantitative results show that the addition of VIB constraint can improve the problem of insufficient priori information of few samples to a certain extent and alleviate the influence of various types of noise in high-resolution remote sensing images on the change detection results. And it further confirms that it is reasonable and effective for us to introduce VIB into the field of high-resolution remote sensing images change detection.

E. Effect of the Amount of Fine-Tuning Data

Since the number of labeled samples is one of the keys affecting the performance of deep learning methods, this section is mainly used to analyze the relationship between the performance of different methods and the number of labeled samples. For this reason, 1%, 10%, and 20% labeled samples are randomly selected from all samples in both datasets and the change detection network is fine-tuned. In this case, all the methods adopt the same pretrained model in the experiments with different data amounts. Since the Multiview does not require sample training for change detection, the experimental results remain the same for different data amounts. The experimental results are shown in Fig. 6.

Fig. 6(a) and (b), respectively, shows the change detection performance of different methods on Lebedev dataset and SenseTime dataset using different numbers of samples. It can be seen that the performance of the ChangeFormer does not improve significantly with the increase in the number of training samples on SenseTime dataset. The main reason for this is that the transformer-based model requires a larger number of samples to take advantage of it. There are only a small number of training samples in the experiment, which limits the model performance. Besides, the performance of different models improves to some extent as the number of training samples increases. This shows that a larger number of samples can provide more sufficient prior knowledge, which makes the trained model have better generalization performance. And among all methods, the improvement of the proposed method is relatively more obvious. It performs well in all four-evaluation metrics and obtains the most advanced performance. In addition, it can be observed that compared with SenseTime dataset, the performance improvement on the Lebedev dataset is more significant and the overall detection effect is better. The reason may be that, first, the total number of samples in Lebedev dataset is much larger than that in SenseTime dataset, so there are relatively more labeled samples used for fine-tuning the network in the above experiment, which further confirms that the network performance is related to the number of samples. Second, the changed objects in SenseTime dataset are transformations between different land types, and some transformations with higher texture and shape similarity increase the difficulty of detection to a certain extent.

F. Effect of the Loss Function

The loss function provides a gradient representation for the network parameter iteration by measuring the difference between the predicted result and ground truth, and back propagation, which is a key to the performance of the model in deep learning. As described in Section III, we incorporate dice loss in the change detection fine-tuning model to reduce the impact of the class imbalance on the model performance. To explore the extent of its impact, we remove the dice loss to conduct comparison experiments on two datasets. In addition, the loss function combining weighted binary cross entropy and dice loss has been shown to be effective in alleviating the class imbalance [63], so it is added to the comparison experiment to further validate the model performance, and the results are given in Table V. Note that we named the model "N_diceloss" and "combined loss," respectively.

As we can see from Table V, the combined loss combining weighted cross entropy and dice loss can improve the overall performance to a certain extent compared to using only binary cross entropy. This demonstrates that improving the class imbalance problem can effectively improve the model performance in the change detection. And in our loss function, the addition of dice loss has greatly improved the detection performance of the model on both datasets, achieving the best performance. Specifically, the recall, F1, and IoU on Lebedev dataset improve

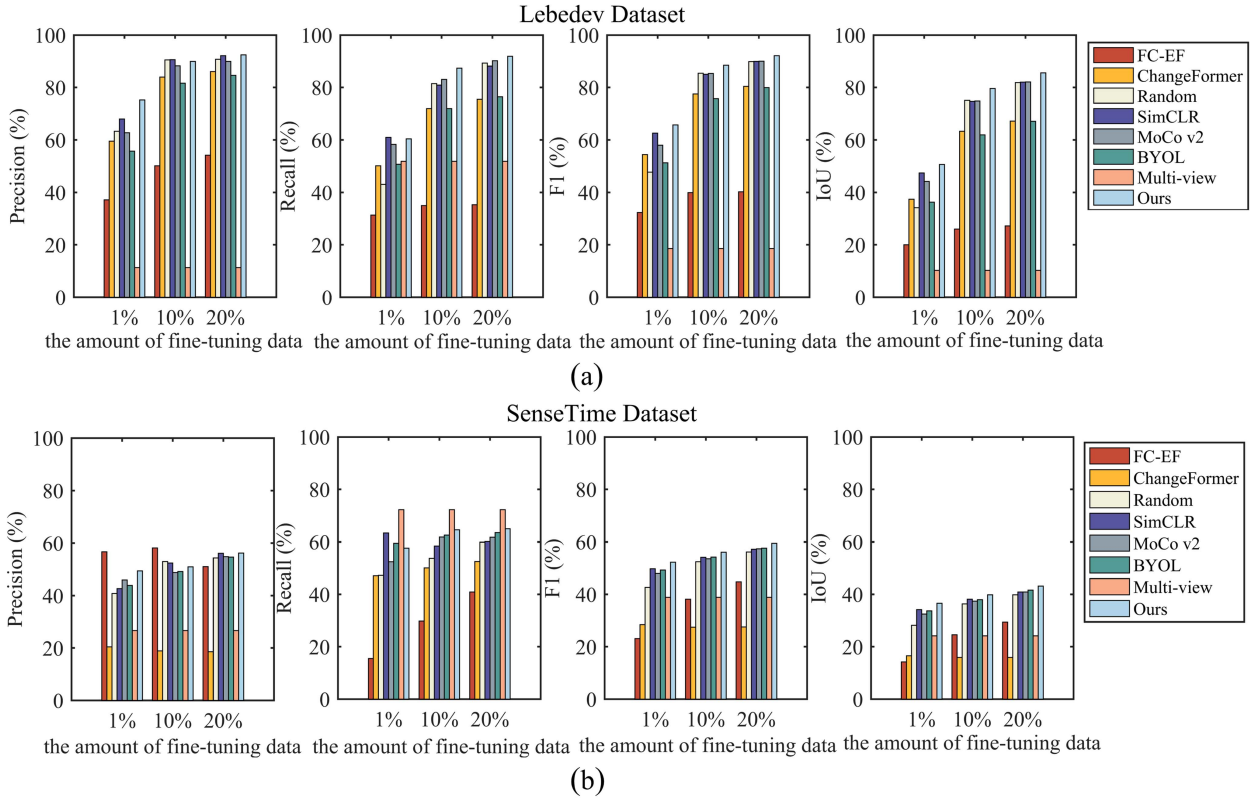

Fig. 6. Change detection performance of different methods with different data amounts on both datasets. (a) Results on the Lebedev dataset. (b) Results on the SenseTime dataset.

TABLE V
EFFECT OF LOSS FUNCTION ON TWO DATASETS

Method	Lebedev Dataset				SenseTime Dataset			
	Precision (%)	Recall (%)	F1 (%)	IoU (%)	Precision (%)	Recall (%)	F1 (%)	IoU (%)
N_diceloss	70.67	59.57	63.13	47.58	51.81	49.37	50.28	33.76
combined loss	72. 18	60.32	64.06	48.86	**56.01**	46.65	50.53	34.27
Ours	**75.23**	**60.44**	**65.79**	**50.61**	49.33	**57.57**	**52.21**	**36.58**

The bolded entities in tables represent the optimal result under the corresponding evaluation indicators.

by 0.87%, 2.66%, and 3.03%, respectively, after the addition of dice loss. Similarly, the improvements on SenseTime dataset are 8.20%, 1.93%, and 2.82%, respectively. This fully demonstrates that it can effectively reduce the missed detection of changed regions, improve the biased prediction caused by the imbalance between the number of changed and unchanged samples to some extent, and improve the overall performance of the model.

G. Effect of the Hyperparameter

1) Effect of β: An important hyperparameter of the proposed method is β in the regularization term of VIB, which controls the tradeoff between information compression and preservation of label-relevant information for the extracted latent features. To investigate its relationship with the detection performance, the model of different β values is evaluated. The results are shown in Fig. 7.

It can be seen from Fig. 7 that there is good detection performance on both datasets when the value β is within a certain range. It benefits from the VIB regularization and makes the model focus on the real change information and filter out part of the noise. However, when β is too large, the VIB regularization focuses on compressing information, resulting in insufficient information retained to identify the changed and unchanged regions, and the model performance decreases. As can be seen from the figure, when β of Lebedev dataset and SenseTime dataset is set to 10^{-7} and 10^{-1}, the extracted features have a good tradeoff between compressing irrelevant information and guaranteeing predictive ability, and the model has optimal performance. In addition, the different optimal settings of β on different datasets also indicate that the effect of β is different

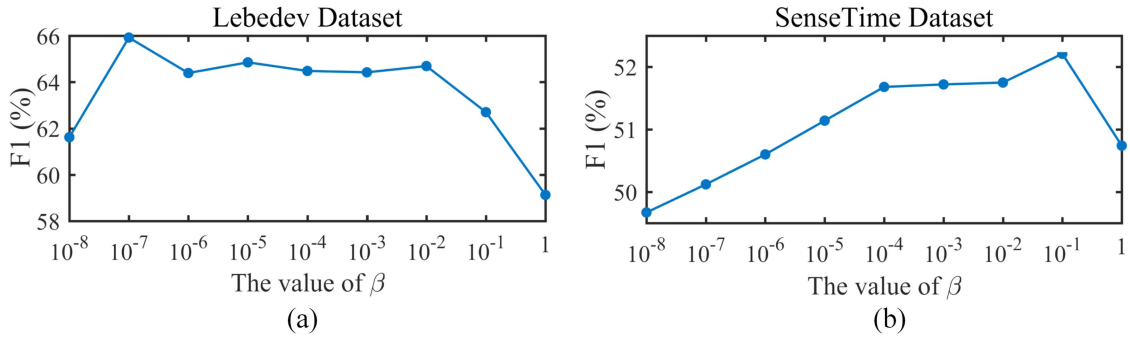

Fig. 7. Change detection performance of different β. (a) Results on the Lebedev dataset. (b) Results on the SenseTime dataset.

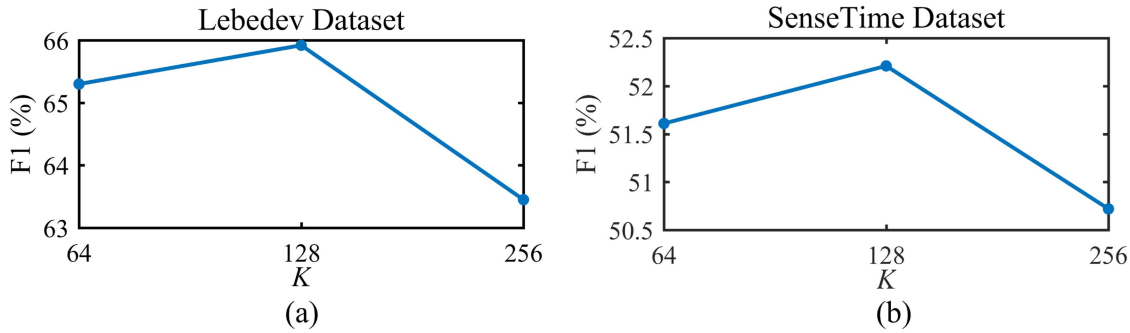

Fig. 8. Change detection performance of different *K*. (a) Results on the Lebedev dataset. (b) Results on the SenseTime dataset.

for different datasets, which may be related to the sample size or the characteristics of the dataset itself.

2) Effect of Feature Dimension K: In addition to β aforementioned, K in (6) of Section III is the dimension of the extracted latent feature, which specifies the bottleneck size. And it works together with β in the tradeoff between information compression and critical information retention. Both increasing β and decreasing K increase the restriction on the flow of feature information. To evaluate the effect of K value on the experimental result, we train the model for varying values of K under the condition that β is fixed as the optimal value of the above-mentioned experimental results on both datasets. And the results are shown in Fig. 8.

When the K value is 128, the model obtains the best overall performance on both datasets. When the K value is smaller, that is, the K value is 64, the model performance is slightly lower than the optimal result. The reason is that the small feature dimension has a stronger restriction on the information transfer, resulting in insufficient ability of the model to encode the feature information, which leads to the decline of the discrimination ability. And the model performance is the weakest when the K value is 256. It is worth noting that in our model, the input feature of VIB module is 256-dimensional. Therefore, setting K to 256 means that the feature information is not compressed, which leads to retaining more irrelevant information, making the overall performance poor. In conclusion, the experimental results further confirm that filtering the feature information to a certain extent can improve the performance of the model. And the effect of the value of K on the results may slightly vary for different model architectures.

V. Discussion

In this work, we apply the contrastive self-supervised mechanism and the theory of VIB to the change detection in high-resolution remote sensing images, and aim to improve model performance with few samples, so as to alleviate the sample dilemma faced by deep learning methods in practical application scenarios and the problem that noise in high-resolution remote sensing images increase the difficulty of change detection. We will further discuss our method performance in this section.

Considering the scarcity of pixel-level labeled samples and the availability of a large number of unlabeled samples in practical applications, and that the self-supervised pretraining method is theoretically capable of learning latent knowledge from unlabeled data, a novel contrastive self-supervised pretraining approach based on this framework is proposed. As expected, our experimental results show that it is difficult for the supervised method to accurately locate regions of change when only a small number of labeled samples are available. However, our change detection model retrained on the basis of pretraining significantly improves the change detection performance, which shows that our method has greater practical value in the case of insufficient labeled samples. This result is in line with the performance of contrastive self-supervised learning in few-sample

TABLE VI
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE LEVIR-CD DATASET

Type	Method	1% of the labeled sample			
		Precision (%)	Recall (%)	F1 (%)	IoU (%)
Supervised method	FC-EF	41.35	40.49	37.44	26.72
	ChangeFormer	**75.97**	46.55	55.43	40.37
Self-supervised method	Random initialization	49.93	38.28	40.74	28.94
	SimCLR	61.41	58.76	58.30	43.31
	MoCo v2	61.47	55.26	56.17	41.23
	BYOL	58.53	60.66	57.48	42.38
	Multiview	1.49	61.37	2.46	1.68
	Ours	64.32	**62.44**	**62.09**	**46.13**

The bolded entities in tables represent the optimal result under the corresponding evaluation indicators.

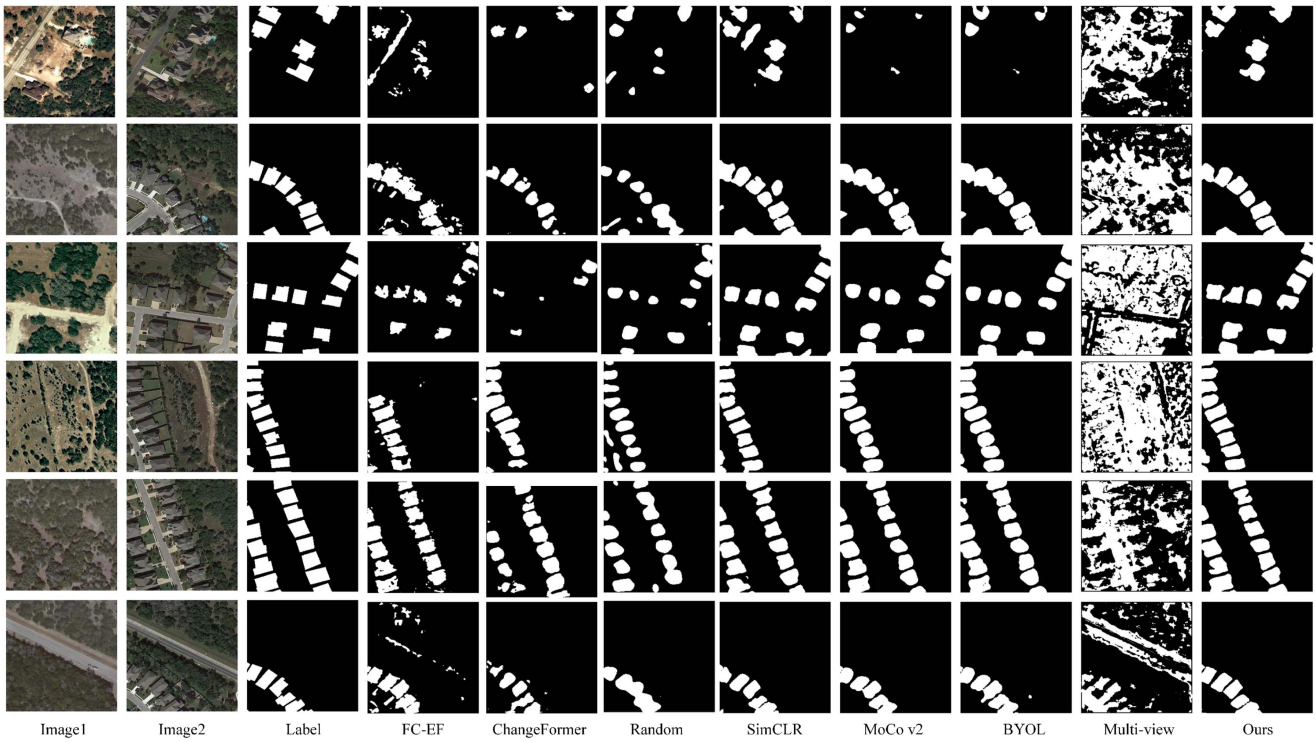

Fig. 9. Qualitative result of different methods on the LEVIR-CD dataset. Black indicates unchanged regions and white indicates changed regions.

image classification [39], segmentation [52], and other fields. And in comparison with the change detection results pretrained by different contrastive self-supervised methods, we find that the performance of different comparative self-supervised network architectures on downstream task varies greatly, with our approach being optimal. The main reason for this is that our proposed contrastive self-supervised model architecture starts from the framework of change detection, thus has a higher adaptability to the actual change detection. In addition, compared with other methods, our proposed multilevel and multigranularity contrastive learning can mine more discriminative information from unlabeled data and can further improve the performance of downstream change detection.

In addition, we find that the proposed VIB regularization constraint has significant advantage in improving the performance of high-resolution images change detection. As shown in the visualization results on both datasets, our model is more robust and the predicted change maps are more refined when

there are disturbances such as shadows and seasonal changes in the image. This fully illustrates the importance of reducing noise information and enhances the attention of the model to the real change information. In the applications of remote sensing, especially the high-resolution remote sensing images, the influence of noise inevitably exists. And the introduction of this theory is of some significance to this situation. Furthermore, the emphasis and attention to feature should be different at different stages of model training. Therefore, greater improvements may be obtained if the hyperparameter β in the VIB regularization term is dynamically adjusted during model training.

To further validate the robustness of our method on different datasets, the proposed model is applied to the LEVIR-CD dataset [64] with change types of building growth and decay, which is a different change detection task for the two datasets above. In the dataset, the total number of labeled sample pairs is 10 192, and the training and test sets in the change detection stage are 1% and 20% of the labeled sample, respectively, which is consistent with the other experimental settings. As can be seen from the quantization results in Table VI, the overall performance of the proposed method is superior to other methods, with precision, recall, F1, and IoU being 64.32%, 62.44%, 62.09%, and 46.13%, respectively. Fig. 9 shows the change results obtained by the proposed model and the comparison methods. As shown in the figure, compared with other methods, the change maps predicted by the proposed method are more consistent with labels and can correctly identify the changed buildings of interest from vegetation and light changes and other pseudochanges. The above-mentioned results show that the proposed method can obtain better performance under the supervision of limited samples in different datasets.

VI. Conclusion

This work focuses on the change detection with a small number of labeled samples and proposes an improved framework of contrastive self-supervised pretraining and a change detection fine-tuning model based on VIB regularization. In particular, the contrastive self-supervised pretraining model contains a four-branch Siamese network feature extractor with multilevel and multigranularity integration, which can learn latent knowledge from a large amount of unlabeled data that can bring gains to downstream task under the constraint of contrastive loss. In addition, the VIB constraint in the stage of fine-tuning change detection explicitly adds prior knowledge of suppression noise, emphasizing change to the model. We performed pretraining and change detection on three public datasets, and the experimental results demonstrate the effectiveness of the proposed method. Specifically, in the quantitative experiments on 1% labeled samples of Lebedev dataset, SenseTime dataset, and LEVIR-CD dataset, the comprehensive evaluation index F1 and IoU of our method achieve 65.79%, 52.21%, 62.09% and 50.61%, 36.58%, 46.13%, respectively, which is superior to other comparison methods. The proposed contrastive learning framework can make the extracted features more discriminative, thus providing a better initial optimization direction for the downstream change detection. And the VIB regularization constraint can reduce the noise information in the high-resolution images to a certain extent and further improve the change detection performance. In the future, we will further investigate semantic change detection that can reveal semantic categories before and after the change for more fine-grained analysis.

References

[1] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using tri-temporal logic-verified change vector analysis," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 278–293, 2020.

[2] Z. Du, J. Yang, C. Ou, and T. Zhang, "Agricultural land abandonment and retirement mapping in the Northern China crop-pasture band using temporal consistency check and trajectory-based change detection approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4406712.

[3] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412712.

[4] A. Vetrivel, M. Gerke, N. Kerle, F. Nex, and G. Vosselman, "Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 45–59, 2018.

[5] P. Ye, G. Liu, and Y. Huang, "Geographic scene understanding of high-spatial-resolution remote sensing images: Methodological trends and current challenges," *Appl. Sci.*, vol. 12, no. 12, 2022, Art. no. 6000.

[6] 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.

[7] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[8] S. Ji, D. Wang, and M. Luo, "Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3816–3828, May 2021.

[9] B. Fang, G. Chen, G. Ouyang, J. Chen, R. Kou, and L. Wang, "Content-invariant dual learning for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603317, doi: 10.1109/TGRS.2021.3064501.

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

[11] C. Connors and R. R. Vatsavai, "Semi-supervised deep generative models for change detection in very high resolution imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 1063–1066.

[12] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.

[13] 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.

[14] Z.-G. Liu, Z.-W. Zhang, Q. Pan, and L.-B. Ning, "Unsupervised change detection from heterogeneous data based on image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403413.

[15] G. Yang, H.-C. Li, W.-Y. Wang, W. Yang, and W. J. Emery, "Unsupervised change detection based on a unified framework for weighted collaborative representation with RDDL and fuzzy clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8890–8903, Nov. 2019.

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

[17] Y. Chen and L. Bruzzone, "Self-supervised change detection in multiview remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5402812, doi: 10.1109/TGRS.2021.3089453.

[18] X. Jiang, G. Li, X.-P. Zhang, and Y. He, "A semisupervised siamese network for efficient change detection in heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4700718.

[19] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4258–4267.

[20] 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.*, 2020, pp. 9729–9735.
[21] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 21271–21284.
[22] 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.
[23] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.
[24] Z. Cai, Z. Jiang, and Y. Yuan, "Task-related self-supervised learning for remote sensing image change detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 1535–1539.
[25] H. Lee and H. Kwon, "Self-supervised contrastive learning for cross-domain hyperspectral image representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 3239–3243.
[26] P. Arsomngern, C. Long, S. Suwajanakorn, and S. Nutanong, "Self-supervised deep metric learning for pointsets," in *Proc. IEEE 37th Int. Conf. Data Eng.*, 2021, pp. 2171–2176.
[27] Y. Fabel et al., "Applying self-supervised learning for semantic cloud segmentation of all-sky images," *Atmospheric Meas. Techn.*, vol. 15, no. 3, pp. 797–809, 2022.
[28] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.
[29] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018.
[30] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1422–1430.
[31] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.
[32] H. Zhong et al., "A self-supervised learning based framework for automatic heart failure classification on cine cardiac magnetic resonance image," in *Proc. IEEE 43rd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2021, pp. 2887–2890.
[33] G. Díaz, B. Peralta, L. Caro, and O. Nicolis, "Co-training for visual object recognition based on self-supervised models using a cross-entropy regularization," *Entropy*, vol. 23, no. 4, 2021, Art. no. 423.
[34] Z. Yang, H. Yu, Y. He, W. Sun, Z.-H. Mao, and A. Mian, "Fully convolutional network-based self-supervised learning for semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2022.3172423.
[35] B. Dang and Y. Li, "MSResNet: Multiscale residual network via self-supervised learning for water-body detection in remote sensing imagery," *Remote Sens. (Basel)*, vol. 13, no. 16, 2021, Art. no. 3122.
[36] D. Yuan, X. Chang, P.-Y. Huang, Q. Liu, and Z. He, "Self-supervised deep correlation tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2021.
[37] Y. Chen, F. Wu, Z. Wang, Y. Song, Y. Ling, and L. Bao, "Self-supervised learning of detailed 3d face reconstruction," *IEEE Trans. Image Process.*, vol. 29, pp. 8696–8705, 2020.
[38] R. Hang, X. Qian, and Q. Liu, "Cross-modality contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532812, doi: 10.1109/TGRS.2022.3188529.
[39] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9588–9597.
[40] 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.
[41] N. Tishby, C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Allerton Conf. Commun. Control Comput.*, Jul. 2001, vol. 49.
[42] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2016.
[43] Q. Lai, Y. Li, A. Zeng, M. Liu, H. Sun, and Q. Xu, "Information bottleneck approach to spatial attention learning," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Int. Joint Conf. Artif. Intell. Org., Z.-H. Zhou, Jun. 2021, pp. 779–785, doi: 10.24963/ijcai.2021/108.
[44] A. Zhmoginov, I. Fischer, and M. Sandler, "Information-bottleneck approach to salient region discovery," in *Proc. Mach. Learn. Knowl. Discov. Databases, Eur. Conf.*, 2021, pp. 531–546.
[45] J. Kim and M. Bansal, "Towards an interpretable deep driving network by attentional bottleneck," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 7349–7356, Oct. 2021.
[46] D. Arumugam and B. Van Roy, "Deciding what to learn: A rate-distortion approach," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 373–382.
[47] C. Bai et al., "Dynamic bottleneck for robust self-supervised exploration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 17007–17020.
[48] Y. Belinkov, and J. Henderson, "Variational information bottleneck for effective low-resource fine-tuning," in *Proc. Int. Conf. Learn. Representations*, 2020.
[49] Z. Wan, C. Zhang, P. Zhu, and Q. Hu, "Multi-view information-bottleneck representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10085–10092.
[50] Q. Sun et al., "Graph structure learning with variational information bottleneck," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 4165–4174.
[51] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *Proc. Int. Conf. Learn. Representations*, 2020.
[52] 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.
[53] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *Adv. in Neural Inf. Process. Syst.*, 2018.
[54] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.
[55] Y. Zhou, W. Huang, P. Dong, Y. Xia, and S. Wang, "D-UNet: A dimension-fusion U shape network for chronic stroke lesion segmentation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 940–950, May/Jun. 2021.
[56] C. Wang, W. Sun, D. Fan, X. Liu, and Z. Zhang, "Adaptive feature weighted fusion nested U-net with discrete wavelet transform for change detection of high-resolution remote sensing images," *Remote Sens. (Basel)*, vol. 13, no. 24, 2021, Art. no. 4971.
[57] 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. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, 2018.
[58] Sense Time, "Sense Earth 2020-change detection," 2020. [Online]. Available: https://aistudio.baidu.com/aistudio/datasetdetail/53484
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[60] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.
[61] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[62] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.
[63] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8017505.
[64] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens. (Basel)*, vol. 12, no. 10, 2020, Art. no. 1662.

Congcong Wang is currently working toward the doctor's degree in geodesy and survey engineering with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing, China.

Her research interests include remote sensing images change detection and deep learning.

Shouhang Du received the Ph.D. degree in cartography and geographic information system from Peking University, Beijing, China, in 2021.

He is currently a Lecturer with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing. His research interests include intelligent understanding of spatial data, GIS, and remote sensing data.

Wenbin Sun received the Ph.D. degree in cartography and geographical information engineering from the China University of Mining and Technology, Beijing, China, in 2007.

He is currently with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing, China, as a Professor and a Doctoral Supervisor. His research interests include global discrete grid theory and methods, Big Data analysis application, intelligent computing, and deep learning.

Deqin Fan received the Ph.D. degree in cartography and geographic information system from Beijing Normal University, Beijing, China, in 2014.

She is currently with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing, as an Assistant Professor. Her research interests include remote sensing data processing, vegetation and ecology remote sensing, and spatiotemporal variation of vegetation phenology.