Semantic-Aware Dense Representation Learning for Remote Sensing Image Change Detection

Hao Chen, Graduate Student Member, IEEE, Wenyuan Li, Song Chen, and Zhenwei Shi, Member, IEEE

Abstract—Supervised deep learning models depend on massive labeled data. Unfortunately, it is time-consuming and labor-intensive to collect and annotate bitemporal samples containing desired changes. Transfer learning from pretrained models is effective to alleviate label insufficiency in remote sensing (RS) change detection (CD). We explore the use of semantic information during pretraining. Different from traditional supervised pretraining that learns the mapping from image to label, we incorporate semantic supervision into the self-supervised learning (SSL) framework. Typically, multiple objects of interest (e.g., buildings) are distributed in various locations in an uncurated RS image. Instead of manipulating image-level representations via global pooling, we introduce point-level supervision on per-pixel embeddings to learn spatially sensitive features, thus benefiting downstream dense CD. To achieve this, we obtain multiple points via class-balanced sampling on the overlapped area between views using the semantic mask. We learn an embedding space where background and foreground points are pushed apart, and spatially aligned points across views are pulled together. Our intuition is the resulting semantically discriminative representations invariant to irrelevant changes (illumination and unconcerned land covers) may help change recognition. We collect large-scale image-mask pairs freely available in the RS community for pretraining. Extensive experiments on three CD datasets verify the effectiveness of our method. Ours significantly outperforms ImageNet pretraining, in-domain supervision, and several SSL methods. Empirical results indicate our pretraining improves the generalization and data efficiency of the CD model. Notably, we achieve competitive results using 20% training data than baseline (random initialization) using 100% data. Our code is available at https://github.com/justchenhao/SaDL_CD.

Index Terms—Change detection (CD), convolutional neural networks (CNNs), remote sensing (RS) image, self-supervised learning (SSL), semantic-aware representation learning.

I. INTRODUCTION

REMOTE sensing (RS) image change detection (CD) is the process of identifying changes of interest in a scene based on multitemporal RS images captured at different times [1]. The definition of change varies across applications, such as urban expansion [2], agricultural surveys [3], deforestation monitoring [4], disaster damage assessment [5], etc. The available multitemporal very high-resolution (VHR) RS images from satellite and aerial data provide possibilities for monitoring land cover changes at a fine scale. CD based on VHR images remains challenging due to two aspects: 1) the complexity of objects in a scene, i.e., objects of the same semantic category may exhibit varying spectral properties in different geospatial locations and seasons and 2) diverse imaging conditions, e.g., sensor characteristics and illumination variances. [6]. A good CD model can identify real changes (e.g., buildings) while ignoring irrelevant changes (e.g., illumination and unconcerned land-cover changes).

Deep learning (DL) techniques have been widely applied in the RS CD task [7]. The success of DL-based supervised models heavily relies on large labeled data. Unfortunately, it is time-consuming and labor-intensive to collect and annotate bitemporal samples that contain changes of interest. Transfer learning is an effective solution to handle the CD labeled data insufficiency. Its basic concept is to leverage the knowledge of a pretrained model previously trained on a large-scale dataset onto a downstream task that contains relatively small labeled data. ImageNet pretraining is commonly applied in contemporary CD models [8], [9], [10], [11], [12], [13]. Considering the domain gap between natural and RS images may induce suboptimal representations, a new trend is to pretrain on RS data to learn in-domain representations, including supervised RS pretraining [10], [14], [15] and unsupervised RS pretraining based on self-supervised learning (SSL) [16], [17], [18], [19]. In this work, we focus on supervised RS pretraining via leveraging existing RS data in the RS community. Different from existing supervised pretraining approaches that learn the mapping from the image to the label, we utilize the semantic supervision in a contrastive manner via incorporating it into the SSL framework.

Contrastive SSL [20], [21] could learn meaningful representations from massive unlabeled data by pulling together representations of semantically similar samples (i.e., positive pairs) and pushing apart those of dissimilar samples (i.e., negative pairs). Very recently, contrastive methods have been introduced in the RS domain [16], [17], [18], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33] and have shown promising performance for the downstream supervised CD task [16], [17], [18].
Most existing contrastive methods in CD express each image view into one representation vector via uniform aggregation of pixel-level embeddings. Different from the object-centered natural image (e.g., ImageNet), typically multiple small objects (e.g., buildings) are located in various positions in a real-world RS image. Contrastive learning on global representations that lack spatial information is not optimal for the downstream CD dense prediction task. Moreover, land cover spatial misalignment between different views (i.e., positive pairs) of the same image via artificial augmentations (e.g., random crop) may further hinder representation learning.

Based on the above observations, we propose semantic-aware dense representation learning for RS image CD. Fig. 1 illustrates the proposed method. Instead of manipulating global representations, we constrain dense consistencies of the per-pixel image embeddings according to the location correspondence in the overlapped area between the two views. Multiple spatially aligned points across views are sampled to achieve dense correspondence. Our motivation is that the introduced point consistency may not only learn the high-level representations invariant to different augmentations (i.e., color and geometry transformations) but also sustain sensitivity to local spatial information, which helps the downstream dense prediction task. Toward semantic-aware representation learning, we obtain class-balanced sample points from each semantic region by leveraging the semantic mask. We follow SimSiam [21] to implement the cross-view (i.e., a positive pair) similarity of the spatially aligned points. We additionally employ a semantic dissimilar loss to push away foreground and background pixels in the embedding space. Such semantic discrimination (SD) design helps the downstream model to identify objects of interest (i.e., foreground). Apart from two views by basic augmentations to the image, we generate the third view via background swap (BS) to achieve consistent foreground representations invariant to irrelevant background changes. The synthetic view of the current image is generated by blending its foreground and the background from another image using the semantic mask.

We collect a pretraining dataset that contains large-scale RS images and corresponding building semantic masks. Notably, we leverage image-mask pairs freely available in RS society [34]. Our proposed method is evaluated on three public CD datasets [2], [35], [36]. Extensive experiments on the downstream CD task verify the effectiveness of the proposed pretraining.

The contribution of our work can be summarized as follows.

1) Different from traditional supervised pretraining that learns the mapping from the image to the label, we explore the use of semantic information in a representation learning framework and propose semantic-aware pretraining based on class-balanced sampling for RS image CD. Instead of manipulating image-level representations, we constrain pixel-level cross-view consistency as well as SD to learn spatial sensitive features, thus benefiting the downstream dense prediction CD task.

2) Different from the self-supervised baseline, we additionally leverage semantic supervision in a contrastive manner to enhance feature discrimination ability. Apart from two views via basic augmentations, we propose a third view of a BS version to learn representations invariant to irrelevant changes (unconcerned land covers and illumination).

3) Extensive experiments on three CD datasets verify the effectiveness of the proposed method. Ours outperform ImageNet pretraining, in-domain supervision, and several state-of-the-art SSL methods.

The rest of this article is organized as follows. Section II introduces related work of CD data-efficiency approaches and pretraining methods. Section III describes our proposed semantic-aware pretraining method. Experimental results are given in Section IV, and the Conclusion is drawn in Section VI.

II. RELATED WORK

A. Handling Label Insufficiency in CD

Training DL-based models typically rely on a massive amount of labeled data. Despite a large amount of RS data, manually collecting bitemporal images containing real changes and annotating high-quality labels could be costly. Many efforts in RS CD have been made to tackle the label insufficiency, including applying data augmentation [37], [38], [39], [40], [41], [42], [43], generating pseudo labels for unlabeled data via semisupervised learning [36], [44], [45], using active learning to select a small number of informative samples [46], [47], [48], [49], and transfer learning from pretrained models [8], [10], [16], [17], [19], [50].

Data augmentation is an effective solution to enhance the size of the training dataset. The most common way is to use transformation-based augmentations [37], [39], [40], [41], [42], [43], including geometric transformations (e.g., random crop, horizontal flip), color transformations, Gaussian blur, etc. A recent advance increases the number of the positive samples (change of interest) by blending the generative adversarial networks (GAN)-generated instance on the appropriate spatial-temporal position of the bitemporal image [38].

Transfer learning from a pretrained model allows leveraging the knowledge of a source dataset to improve the data efficiency of a downstream task. ImageNet pretraining is widely used in CD [7], [8], [9], [10], [11], [12], [13] and shows superior to random initialization, especially in small data regimes. Considering the domain gap between ImageNet and RS images, a new trend is to pretrain on the RS data to learn the in-domain representations [10], [14], [15], [16], [17], [19], [50], [51], [52], [53]. Supervised RS pretraining normally learns discriminative representations from the RS image-level [10], [15] or pixel-level [14] land-cover classification samples, while self-supervised pretraining could learn meaningful representations by making use of a vast amount of unlabeled RS data [16], [17], [19].

Our method falls in the supervised pretraining scope. Different from normal supervised pretraining methods that learn the mapping from image to label, we incorporate the semantic
supervision in a contrastive manner in the SSL framework to enhance the feature discrimination ability. Notably, we leverage image-mask pairs freely available in RS society to guide the contrastive pretraining.

Please note that a shorter, preliminary study of our semantic-aware representation learning has been presented in the International Geoscience and Remote Sensing Symposium (IGRASS) 2022 [53]. This article is a continuation and extension of our previous work by introducing two more pretraining designs related to the use of semantic information as well as supplementing more extensive experimental results, including more comparisons, ablations, and visualizations.

B. Contrastive Pretraining Methods

SSL has attracted many researchers by virtue of its ability to learn good feature representations from massive unlabeled data to benefit various downstream tasks [54]. Currently, contrastive methods (e.g., a Simple framework for Contrastive Learning of visual Representations (SimCLR) [55], Momentum Contrast for unsupervised visual representation learning (MoCo) [20, 56], Bootstrap Your Own Latent (BYOL) [57], SimSiam [21], and dense contrastive learning (DenseCL) [58]) have made significant progress in SSL. The key to contrastive approaches is to perform the instance discrimination pretext task by pulling together representations of similar images (positive pairs) and pushing apart those of dissimilar images (negative pairs). The InfoNCE loss [59] is typically applied to induce similarity between positive pairs and dissimilarity between negative pairs [11, 20, 55]. More recent attempts [21, 57, 60] apply simpler loss functions (e.g., negative cosine similarity and cross-entropy) that only rely on positive samples and does not use negative ones.

Very recently, contrastive SSL methods have been introduced in RS to obtain in-domain representations. Contrastive pretraining have been shown to benefit RS downstream tasks, including land-cover classification [22, 24, 25, 26, 27, 28, 61, 62], semantic segmentation [29, 30, 31, 32], and CD [16, 17, 18]. Most existing methods apply InfoNCE loss [18, 22, 24, 25, 26, 27, 28, 30, 33, 63] or triple loss [17, 61] on the constructed positive and negative pairs. Positive samples can be obtained by different artificial augmentations (e.g., color and geometric transformations) of the same image [25, 28], spatial augmentations (i.e., geospatially overlapped images) [24, 27, 61], temporal augmentations (i.e., multitemporal coregistered images) [16, 17, 22], [26, 63], and modality augmentations (e.g., optical image, synthetic aperture radar (SAR), and semantic mask) [29, 64]. Negative pairs can be different samples in a mini-batch or spatially distinct images [61, 63].

Contrastive methods for CD have seldom been explored. Some attempts apply SSL directly on a small downstream CD dataset to extract seasonal invariant features for unsupervised CD [65, 66, 67]. Differently, we employ an additional relatively large-scale RS dataset to learn semantic-aware representations and then transfer knowledge from the pretrained model to the downstream supervised CD task on a variety of CD datasets. Other more related studies that follow the normal SSL pipeline mostly evaluate the pretrained model on the medium-resolution CD dataset [16, 17, 18]. We instead explore pretrained models suitable for high-resolution RS image CD. Instead of manipulating global representations, we propose dense points suitable for high-resolution RS image CD. Instead of manipulating global representations, we propose dense points suitable for high-resolution RS image CD. Instead of manipulating global representations, we propose dense points suitable for high-resolution RS image CD. Instead of manipulating global representations, we propose dense points suitable for high-resolution RS image CD. Instead of manipulating global representations, we propose dense points suitable for high-resolution RS image CD.
RS images that contain plenty of spatial details. Instead of manipulating image-level representations, we constrain point-level consistencies according to spatial correspondences of the sampled pixels across positive views to better sustain the sensitivity of the spatial information, therefore benefiting the downstream dense prediction task, especially on VHR RS images that contain plenty of spatial details. Unlike the existing dense SSL method (DenseCL) that calculates similarities between representations of any two pixels to estimate pixel correspondence across views, our design utilizes the exact pixel correspondence calculated according to known geometric relations between views for learning view-invariant representations. Moreover, we leverage the semantic information in a contrastive manner to further improve the model discrimination ability to distinguish land covers of interest and other unconcerned backgrounds. Most existing methods [16], [17], [65], [66], [67] treat coregistered multitemporal images as positive samples to learn the seasonal invariant features. However, the intrinsic change of interest within the bitemporal image may induce the pretrained model invariant to real changes, which may be unfavorable to the downstream CD task. Different from existing methods using natural bitemporal images, ours learn consistent foreground representations between one image and its synthetic version whose background is replaced by that from another image. In this way, we leverage synthetic bitemporal images to enable the model to focus more on foreground objects regardless of changes in unconcerned land covers.

C. DL-Based Models in RS

The recent few years witness the great success of DL techniques, e.g., convolutional neural networks (CNNs), graph convolutional networks (GCNs) [68], and transformers [69], applied in various RS tasks [70], including land cover classification/segmentation [68], [69], multimodal image classification [71], [72], object detection [73], and CD [2], [7], [37], [38], [39], [40], [41], [42], [43], [74], [75], [76], [77]. CNNs have been demonstrated to be effective to extract meaningful and transferable image representations by multiple levels of abstraction, thus attracting much attention in RS society. Most of the existing DL-based CD models [2], [37], [38], [39], [40], [41], [42], [43] are based on CNNs. GCNs originally designed for graph data and transformers for sequential data have also been employed in RS image understanding tasks [68], [69], including CD [74], [75], [76], [77], for improving model discriminative ability by leveraging their capacity of modeling long-range relations among image pixels. In this article, instead of exploring sophisticated model structures for the CD task, we focus on the representation learning method to improve the downstream conventional CNN-based CD model.

III. SEMANTIC-AWARE DENSE REPRESENTATION LEARNING

In this section, we first give an overview of the proposed semantic-aware pretraining for the downstream CD task and then introduce its three main components. Finally, implementation details are given.

A. Overview

To alleviate the insufficiency of labeled CD data, we explore a supervised representation learning method by leveraging massively available data in the RS society. We collect large-scale image-mask pairs (see Section IV-A) in a painless manner for contrastive pretraining. Specially, we focus on the building category as the objects of interest to perform CD.

We propose Semantic-aware Dense representation Learning (SaDL) for RS image CD. The goal of SaDL is to learn a feature extractor that could: 1) generate consistent representations invariant to illumination changes as well as changes of unconcerned land covers (i.e., background); 2) distinguish objects of interest from the background; and 3) sustain spatially sensitive features to reduce the gap with the downstream dense prediction CD task.

To achieve this, we introduce the semantic mask into a SSL framework. Fig. 1 illustrates the proposed SaDL. Following SimSiam [21], we use simple Siamese networks as the feature extractor to learn meaningful image representations without using negative samples. Our SaDL is composed of three main steps.

1) Views Generation: Apart from two views of the same image via normal augmentations (i.e., color and geometry transformations), we generate a third view by swapping its background with that from another image using the semantic mask. These views can be used to learn consistent representations invariant to illumination and background changes. More generation details are given in Section III-B.

2) Masked Sampling: Instead of obtaining image-level representations, our Siamese encoder converts each image view into dense per-pixel representations. We sample the same number of feature points from each semantic region in the overlapped area of the generated views. For more details, see Section III-C.

3) Representation Learning: Given the sampled point representations, we employ multilayer perceptron (MLP)-based projection/prediction networks to perform feature transformations. We constrain point-level cross-view similarity and semantic dissimilarity to learn representations that are invariant to irrelevant changes and discriminative to foreground/background. For more details, see Section III-D.

B. Views Generation

Given each input training sample (including an image I and a mask M) in a mini-batch, we produce three views of the same image, including two basic transformed versions, and one BS version. Let \( T = T_s \circ T_c \) denotes a set of artificial augmentations, including color augmentations \( T_c \) (e.g., color jetting and Gaussian blur) and geometric augmentations \( T_g \) (e.g., random cropping and flipping). Details of our views generations are described in Algorithm 1.

We first generate two views \( I^1, i \in \{1, 2\} \) of I via different artificial augmentations

\[
I^1 = T(I) = T_g(T_c(I))
\]
where $\mathcal{T}, \mathcal{T}'$ denote two random sets of artificial augmentations.

The corresponding semantic masks $M^i$ for each view $i \in \{1, 2\}$ are generated via applying the same geometric augmentations as that in the image

$$M^1 = \mathcal{T}_g(M)$$
$$M^2 = \mathcal{T}'_g(M).$$  \hfill (2)

1) **View Generation Via BS:** The third view of $I$ is obtained by swapping the background of the first view $I^1$ with that of another image $I'$. For simplicity, $I'$ has the inverted index as $I$ from the same mini-batch. In other words, the summation of the index of $I$ and that of $I'$ equals the mini-batch size. Note that we swap the common background area between $I^1$ and $I'$ to simulate the changes of unconcerned land covers. We obtain the common background mask $M_{bg}$ via calculating the intersection of the background regions in $M^1$ and $M^2$.

However, directly pasting the background of $I'$ on $I^1$ using $M_{bg}$ may result in undesired synthetic artifacts [see Fig. 2(d)] that could be detrimental to representation learning. Toward a realistic and effective image composite, we present a simple view synthesis method, including two main steps.

1) **Color Transfer:** To reduce the domain gap between $I'$ (source image) and $I^1$ (target image), we employ a simple color transfer method [78] to change the color characteristic of one image to accord with another in the image domain. The key idea is to match the mean and variance of each channel of the transformed image to that of the target image. To be formal, the transformed image is given by

$$I'^i = \sigma(I^1) \left( \frac{I'^i - \mu(I'^i)}{\sigma(I'^i)} \right) + \mu(I^1)$$  \hfill (3)

where $\mu(\cdot), \sigma(\cdot)$ denote the means and variances computed across spatial dimensions independently for each channel of the image.

2) **Context-aware blending:** The mere presence of the context surrounding the foreground (e.g., building) is a critical cue for object recognition. For a realistic composite, we preserve pixels near the foreground in image $I^1$ by performing morphological erosion on the background mask. We further blur the eroded background mask using a Gaussian filter toward natural and seamless composition. The composite image $I^3$ (view 3) is calculated by alpha blending using the eroded mask of the blurred version

$$I^3 = (1 - M_{bg})I^1 + M_{bg}I'^1.$$  \hfill (4)

Fig. 2 illustrates an example of the third view synthesis. We can observe that using color transfer and context-aware blending could give a more realistic image composite.

C. **Masked Sampling**

1) **Obtaining Dense Representations:** The generated three views $I^1, I^2, I^3$ are then encoded by a Siamese neural network $f$ into dense representations $X^i = f(I^i) \in \mathbb{R}^{H \times W \times C}$, $i \in \{1, 2, 3\}$. The resulting feature map has a size of $H \times W$ and a channel dimension of $C$. Our encoder has a downsampling factor of 4. For more details, see Section III-E.

Please note that we do not apply global average pooling to per-pixel image embeddings. Instead of manipulating image-level representations, we perform dense point-level supervision via sampling multiple points in the overlapped area across views. Our intuition is that the learned spatially sensitive
Algorithm 1 Views Generation for a Batch of Samples

Input: \( \hat{I} \in \mathbb{R}^{H \times W \times 3} \) (a batch of input images)
Input: \( M \in \mathbb{R}^{H \times W} \) (a batch of corresponding semantic masks)
Output: \( \hat{I}^1, \hat{I}^2, \hat{I}^3 \in \mathbb{R}^{H \times W \times 3} \) (a batch of image view 1/2/3)

1 // views generation for each sample in the batch
2 for \( b \) in 1 : \( B \) do
3     // generate view 1 and view 2 of the current image
4     \( \hat{I}^1[b], \hat{I}^2[b] \leftarrow T(\hat{I}[b]), T'(\hat{I}[b]) \)
5     \( M^1[b], \hat{I}^1[b] \leftarrow T_g(M[b]), T'_g(M[b]) \)
6 end
7 for \( b \) in 1 : \( B \) do
8     // generate view 3 of the current image
9     set the sample of an inverted index as the background image/mask
10    \( \hat{I}' \leftarrow \hat{I}^1[B - b] \)
11    \( M' \leftarrow \hat{M}^1[B - b] \)
12    \( \hat{I}' \leftarrow \text{ColorTransfer}(\hat{I}', \hat{I}^1[b]) \)
13     // obtain the common background mask
14    \( M_{bg} \leftarrow (\hat{M}^1[b] == 0) \cdot (\hat{M}' == 0) \)
15     // context-aware blending
16    \( M_{bg} \leftarrow \text{Erosion}(M_{bg}) \)
17    \( \hat{I}[b] \leftarrow (1 - M_{bg}) \hat{I}^1 + M_{bg} \hat{I}' \)
18 end

features may better reduce the gap between the pretraining and downstream dense prediction CD task that requires spatially variant representations. Specifically, we introduce the semantic mask to achieve class-balanced sampling, i.e., an equal number of points is randomly sampled for each semantic region in the overlapped area. The details of our dense representation learning based on masked sampling on the generated views for each sample are described in Algorithm 2.

2) Overlapped Region Between Views: Different views (view 1 and view 2) may correspond to different spatial areas in the original image due to the difference in applied geometry augmentations. To perform cross-view point consistency, we first need to obtain the overlapped region between \( \hat{I'}^1 \) and \( \hat{I'}^2 \). Given the known geometry augmentations (\( T_g \) and \( T'_g \)), we could calculate the corresponding bounding box region (\( BB' = (u', v', w', h') \)) of view \( i, j \in \{1, 2\} \) in the original image \( I \) by utilizing the parameters of each geometry augmentation (e.g., random cropping and random flipping) in the inverse order. For example, random cropping gives the upper-left coordinate \( (u, v) \), the width \( w \), and the height \( h \) of the cropped area in the input image. The overlapped region \( BB \) is the intersection of the two bounding box regions \( BB^1 \) and \( BB^2 \). The region within \( BB \) defines the valid sampling region and other regions outside \( BB \) are invalid for the latter point sampling process.

Please note that we do not consider view 3 for calculating the overlapping region because we only take care of the foreground of view 3, which is shared with view 1.

Fig. 3(a) gives an example of visualization of the view correspondence. View 1 (\( \hat{I}^1 \)) and view 2 (\( \hat{I}^2 \)), respectively,
correspond to the region within the green/yellow bounding box in the original image \( I \). The overlapped area between views in form of bounding boxes (red) is shown in Fig. 3(b).

3) Point Sampling in the Overlapped Region: We sample \( N \) points \( P_{kn} = (u_{kn}, v_{kn}) \) with replacement for each semantic category \( k \) in the overlapping region, where \( (u_{kn}, v_{kn}) \) denotes the upper-left coordinate of point \( P_{kn} \) in the original image/mask \( (I/M) \). \( N \) is the number of sampled points for each class. Our default \( N \) is set to 16 (See Section III-E). \( k = 1 \) denotes the background and \( k = 2 \) denotes the foreground class. Let \( P = \{ P_{kn} = (u_{kn}, v_{kn}) \} \in \{1, 2, \ldots, N\}, k \in \{1, 2\} \) denotes the sampled point set. The corresponding points in view 1 and view 2 can be obtained by coordinate transformation under the given geometry augmentations

\[
P^1 = T_3(P) \quad \text{and} \quad P^2 = T_2(P)
\]

where \( P^i = \{ P^i_{kn} = (u^i_{kn}, v^i_{kn}) \} \in \{1, 2, \ldots, N\}, k \in \{1, 2\} \) denotes the set of coordinates of the sampled point in view \( i, i \in \{1, 2\} \). Note that we divide \( P^i \) by \( d \), i.e., the downsampling factor of the encoder, defaults to 4 to match the coordinates of the per-pixel representations \( X^i \). The sampled points set for view 3 is given by \( P^3 = \{ P^3_{kn} = (u^3_{kn}, v^3_{kn}) \} \in \{1, 2, \ldots, N\}, k = 2 \).

Then for each view \( i, i \in \{1, 2\} \), we gather \( N \) representations \( x^i_{kn} \) in the embedding space \( V \subset \mathbb{R}^C \) for each semantic category \( k \) by indexing points on the feature map \( X^i \) using coordinates \( P^i \)

\[
x^i = \text{gather}(X^i, P^i)
\]

where \( x^i = \{ x^i_{kn} \} \in \{1, 2, \ldots, N\}, k \in \{1, 2\} \) represents the set of representations of sampled points for view \( i \).

Similarly, we gather \( N \) representations \( x^3_{kn} \) for view 3 using sampled points \( P^3 \) that belong to the region of the foreground class \( k = 2 \). Formally, \( x^3 = \{ x^3_{kn} \} \in \{1, 2, \ldots, N\}, k = 2 \).

D. Representation Learning

1) Semantic Discrimination: We exploit the semantic supervision in a contrastive manner via pushing apart representations of points that belong to different semantic regions in the same image. We constrain point-level pair-wise semantic dissimilarity. Our motivation is that the learned discriminative dense features could distinguish the objects of interest and other unconcerned backgrounds, which may facilitate the downstream change recognition of the land covers of interest.

To achieve this, we minimize the cosine similarity between the background and foreground vectors (i.e., point representations from the foreground and the background, respectively) in the embedding space \( V \). For simplicity, we construct \( N \) pairs of points by one-to-one correspondence between \( N \) foreground points \( x^i_{kn} \) and \( N \) background points \( x^i_{kn}, n \in \{1, \ldots, N\} \) for each view \( i \in \{1, 2\} \). Note that we supplement 1 to cosine similarity to ensure a nonnegative loss value. Our semantic dissimilar loss \( \mathcal{L}_{sd,n} \) on one constructed point pair indexed \( n \) for view \( i \) is defined as follows:

\[
\mathcal{L}_{sd,n} = D(x^i_{kn}, x^3_{kn}) + 1
\]

where \( D(a, b) \) denotes the cosine similarity, given by

\[
D(a, b) = \frac{a \cdot b}{\|a\|_2 \cdot \|b\|_2}
\]

where \( \| \cdot \|_2 \) denotes \( l_2 \)-norm, and \( a, b \) are two vectors of the same dimension.

The semantic dissimilar loss for one sample is the average of \( \mathcal{L}_{sd,n} \) on all the constructed point pairs in each view

\[
\mathcal{L}_{sd} = \frac{1}{2N} \sum_{i=1}^{2} \sum_{n=1}^{N} \mathcal{L}_{sd,n}
\]

2) Cross-View Similarity: Given the sampled points from different views, we constrain the pair-wise similarity between point representations by utilizing the spatial correspondences of cross-view points. We aim to learn two kinds of point-level consistencies: 1) consistent representations invariant to illumination changes and 2) consistent foreground representations regardless of background changes. We employ view 1/2 to achieve the first kind consistency and view 1/3 for the second kind.

MLP-based projection/prediction networks are applied to implement the cross-view consistency. For each view \( i \in \{1, 2, 3\} \), the sampled point representations \( x^i_{kn} \) from the encoder are projected into a subspace \( \mathcal{Z} \subset \mathbb{R}^C \) by a projection MLP head (projector) \( g: V \mapsto \mathcal{Z} \) to obtain projection vectors \( z^i_{kn} = g(x^i_{kn}) \). Then, a prediction MLP head (predictor) \( h: \mathcal{Z} \mapsto \mathcal{Z} \) transforms the output of one view (i.e., \( z^i_{kn} \)) to obtain a prediction vector (i.e., \( p^i_{kn} = h(z^i_{kn}) \in \mathbb{R}^C \)) in the same space and matches it to the other view (i.e., \( z^j_{kn} \)). We minimize the negative cosine similarity of representations of corresponding points between the two views \( (i/i') \). We define a symmetrized similarity loss between view \( i \) and view \( i' \) for \( n \)-th point belonging to semantic category \( k \) as follows:

\[
\mathcal{L}_{c,kn}(i/i') = 1 - \frac{1}{2} \left( D(p^i_{kn}, \text{stopgrad}(z^j_{kn})) + D(p^j_{kn}, \text{stopgrad}(z^i_{kn})) \right)
\]

where \( \text{stopgrad} (\cdot) \) denotes the stop-gradient operation. The encoder and projector cannot receive gradient from projection vectors. Previous work [21] has empirically demonstrated the vital importance of such an operation to prevent model collapse.

The first consistency loss is the average similarity loss between view 1 and view 2 on all the sampled points \( n \in \{1, \ldots, N\} \) for each semantic category \( k \in \{1, 2\} \)

\[
\mathcal{L}_{s1} = \frac{1}{2N} \sum_{k=1}^{2} \sum_{n=1}^{N} \mathcal{L}_{c,kn}(1,2)
\]

Similarly, the second consistency loss is the average similarity loss between view 1 and view 3 on all the sampled points \( n \in \{1, \ldots, N\} \) that belong to the foreground \( (k = 2) \)

\[
\mathcal{L}_{s2} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{c,kn}(1,3)
\]
Overall loss function: We calculate the summation of the semantic dissimilar loss and the two kinds of similarity losses to obtain our overall loss for each sample

\[ L = L_{sd} + L_{s1} + L_{s2}. \] (13)

Note that the final loss is averaged over all samples in a mini-batch.

E. Implementation Details

1) Encoder: We use the ResNet-18 [79] backbone (without fully connected layers and global pooling) equipped with feature pyramid networks (FPNs) [80] as our default encoder \( f \) to generate high-resolution and high-level per-pixel features for each image view. The ResNet-18 has 5 stages, each with downsampling by 2. Through the feedforward computation of the backbone, we obtain a feature hierarchy consisting of feature maps at different scales from the activation outputs of the last four stages. The FPN takes as input the feature hierarchy and progressively upsamples the coarse-resolution feature maps while merging them with the corresponding intermediate feature maps by element-wise addition. Each intermediate feature map undergoes a \( 1 \times 1 \) convolutional layer to transform to the same channel dimension \( C \) before addition. The resulting feature map has a 1/4 spatial size of the input image. The channel dimension \( C \) of output features is set to 256.

2) Projector: The projector \( g \) is a 2-layer MLP with a hidden dimension of 2048 and an output dimension of 1024. Batch normalization (BN) and rectified linear unit (ReLU) are added between fully connected (fc) layers.

3) Predictor: The predictor \( h \) is a 2-layer MLP with a hidden dimension of 256 and an output dimension of 1024. BN and ReLU are also added between the two fc layers.

4) Data Augmentation: We set probability of color jittering to 0.8, with \{brightness, contrast, saturation, hue\} strength of \{0.4, 0.4, 0.4, 0.1\}, and the probability of Gaussian blurring to 0.5 with a kernel standard deviation in \[0.1, 2.0\]. Geometric augmentations include random vertical flip, random horizontal flip, and random resized crop with a scale in \[0.8, 1.0\]. Note that we do not follow the common scale [21] for the random crop (i.e., \[0.2, 1.0\]) because the two cropped regions with a small scale may not contain an overlapped area, which may be detrimental to the effective sampling of spatially aligned points across views.

5) Optimizer: We use stochastic gradient descent (SGD) with a momentum hyperparameter of 0.9 and a weight decay of 0.0005. The initial learning rate is 0.01 and decays to zero at the end of training. We follow the “poly” learning rate schedule [81] with a decay coefficient of 0.9. The default number of pretraining epochs \( N_e \) is 200 (see Section IV-E). Due to capacity limitations of memory, the default batch size is set to 64. Our models are implemented on PyTorch and trained using a single NVIDIA RTX 3090 graphics processing unit (GPU).

IV. Experimental Results

A. Pretraining Setup

1) Pretraining Dataset: We leverage image-mask pairs from the existing Inria building segmentation dataset [34], which contains 180 labeled RS images, each size of \( 5000 \times 5000 \) and spatial resolution of 0.3 m. It provides pixel-level annotations, including the building and nonbuilding categories. We cut the original samples into small patches of size \( 256 \times 256 \) with no overlap and remove those without building regions. The resulting dataset contains more than 45 k patch samples. We randomly split it into training (80\%) and validation (20\%) sets. We additionally obtain the coregistered image patch of the corresponding geospatial region via Google Earth as the temporal augmentation \( t_2 \) for each image \( t_1 \) in the dataset. For a fair comparison, the natural augmentations (bitemporal images) are used to train some other related pretraining methods that learn seasonal invariant features. Please note that the semantic mask may not perfectly match the image of \( t_2 \) due to land-cover changes (e.g., building construction and demolition) over time, we do not include the augmented temporal image in our pretraining framework.

B. Downstream CD Setup

1) CD Datasets: To evaluate the proposed pretrained model, we conduct experiments on the following three CD datasets: LEVIR-CD [2]: a widely used building CD dataset. Learning, Vision and Remote sensing (LEVIR)-CD contains 637 pairs of bitemporal VHR RS images, each with a size of \( 1024 \times 1024 \) and a spatial resolution of 0.5 m. We follow the default train-valid-test split [2]. Each image is cropped with no overlap into small patches of size \( 256 \times 256 \). Finally, we obtain 7120/1024/2048 samples for training/validation/testing, respectively.

WHU-CD [35]: a public building CD dataset that includes one pair of optical VHR RS aerial images with a size of \( 32507 \times 15354 \) and 0.075 m spatial resolution. Similarly, we crop the original sample into small patches of size \( 256 \times 256 \) with no overlap. We randomly split the dataset into training, validation, and testing sets, with 6096/762/762 samples, respectively.

GZ-CD [36]: a building CD dataset that covers several suburb areas of Guangzhou (GZ) City, China. It contains 19 VHR red, green, blue (RGB) image pairs (0.55 m) with sizes ranging from \( 1006 \times 1168 \) to \( 4936 \times 5224 \) pixels. Similarly, the image pairs are cropped into \( 256 \times 256 \) nonoverlapping patches. As the dataset provider does not give a default dataset split, we randomly split the dataset into three parts, with \( 2882/361/360 \) samples for training, validation, and testing, respectively.

2) CD Networks: We employ a simple yet effective CD model [38] which consists of a Siamese feature extractor [deep fully convolutional networks (FCN)] to extract high-resolution semantic feature maps for bitemporal patches, a distance metric to calculate the feature difference images (FDI) between the two patches, and a relatively shallow FCN to give the change probability maps. Differently, we employ a light FPN-based decoder head [80] for the feature extractor with the encoder-decoder architecture. Our CD networks employ a ResNet-18 backbone. We use the cross-entropy loss to optimize our CD model. Please refer to [38] for more details.

a) Fine-tuning details: We fine-tune the downstream CD networks using the pretrained models. For a fair comparison,
we only initialize parameters of the backbone (i.e., ResNet-18) of the CD networks by those from the pretrained model. Note that the parameters of other components in the CD networks are initialized with zero-mean normal distribution with a standard deviation of 0.02.

b) Training details: We apply data augmentation, including random flip, and Gaussian blur. Batch size is set to 8. SGD is used for model optimization. We set the momentum to 0.9 and the weight decay to 0.0005. The learning rate starts at 0.01 and linearly decays to 0 until 200 epochs. We perform validation after each training epoch, and use the best model on the validation set for evaluation on the test set.

c) Evaluation Metrics: The F1-score regarding the change category is used as the evaluation indices. Additionally, precision, recall, and Intersection over Union (IoU) of the change category are reported. These metrics are defined as follows:

\[
F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

\[
\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

\[
\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

\[
\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}
\]

where TP, FP, FN denote the number of true positive, false positive, and false negative, respectively.

C. Overall Comparison

To evaluate the effectiveness of the proposed pretraining method, we make a comparison with the random initialization, ImageNet pretraining, in-domain supervised pretraining, and several self-supervised pretraining methods.

1) Random initialization. All the parameters of the CD networks are initialized by the zero-mean normal distribution with a deviation of 0.02.

2) ImageNet pretraining. The backbone (ResNet-18) of the CD model is initialized with the ImageNet pretraining.

3) In-domain supervised pretraining (In-domain Sup.): We employ an FCN-based semantic segmentation network with a ResNet-18 backbone and an FPN head [80], supervised by the image-mask pairs of our pretraining dataset. We follow the same optimization configuration as that for CD networks. We perform the model evaluation in the validation set after each training epoch and save the best model (backbone) for the downstream task.

4) SimSiam [21]: A popular SSL method without using negative samples. SimSiam employs the stop-gradient and predictor to prevent model collapse.

5) MoCo-v2 [56]: A competitive SSL model that formulates contrastive learning as a dictionary look-up problem where a query should be similar to its matching key and dissimilar to others in a dynamic queue covering a large set of negative samples.

6) DenseCL [58]: A DenseCL method that constrains pixel-level cross-view consistencies where the pixel correspondence is computed via Ssing pair-wise similarities between any two pixels from different views.

7) CMC [29]: A recent SSL method in RS takes the semantic mask as an additional view along with the optical/SAR image to learn consistent cross-modal representations under the framework of contrastive multiview coding [82]. Note that we implement multimodal representation learning with image-mask pairs of our pretraining dataset.

8) SeCo [16]: A very recent SSL method that is based on MoCo-v2 and further exploits the use of multitemporal RS images as natural augmentations to learn seasonal invariant/variant representations. We implement it using the bitemporal images in our pretraining dataset.

For a fair comparison, we use our pretraining dataset to train these supervised/self-supervised methods, i.e., in-domain supervision, SimSiam, MoCo-v2, DenseCL, contrastive multiview coding (CMC), and SeCo. The SSL methods are implemented using their public codes with default hyperparameters. Note that all the compared models are initialized with ImageNet pretrained weights as our SaDL does.

Each pretrained model is evaluated by the performance of the downstream CD task. Table I reports the overall comparisons of our proposed method and other pretrained models on the LEVIR-CD, WhuHan University (WHU)-CD, and Guangzhou-CD datasets, respectively. Precision, recall, F1-score, and IoU of the CD networks on the three CD test sets are given. To evaluate the data-efficient performance, we set several data regimes: 1%, 5%, 20%, and 100%, where each number represents the proportion of available labeled data in the original training set.

From Table I, we can observe that the proposed method consistently outperforms the compared pretrained models in terms of the F1-score across different data regimes on the three datasets. Specifically, our SaDL pretraining is significantly superior to random initialization. It is notable that we can achieve comparable or even better results with only 20% training data than the random initialization baseline using 100% training data. It indicates that our pretraining can effectively alleviate the labeled data insufficiency problem.

ImageNet pretraining serves as a strong baseline due to its powerful general representations learned from the large-scale labeled ImageNet dataset. The empirical results demonstrate that our method overachieves ImageNet pretraining across different data conditions on the downstream CD task. For instance, Table I(a) shows that compared to ImageNet pretraining, we achieve considerable improvements of more than 20% of the F1-score in the LEVIR-CD test set under small data regimes (1% and 5%).

1) Comparison to in-Domain Supervision: Typically, supervised RS pretraining obtains in-domain discriminative representations by learning the mapping from RS images to labels. Table I shows the in-domain supervision is consistently inferior to our SaDL under each data configuration of the three datasets, even not comparable to ImageNet pretraining in some data regimes. Note that we use the same amount of pretraining data for SaDL and in-domain supervision. The quantitative results indicate that our SaDL can learn more
TABLE I
OVERALL COMPARISONS OF DIFFERENT PRETRAINING METHODS ON THE THREE CD DATASET. PRECISION/RECALL/F1/IoU OF THE CD MODEL ON THE LEVIR-CD/WHU-CD/Guangzhou-CD TEST SET ARE REPORTED. THE HIGHEST CLASSIFICATION ACCURACY IN EACH DATA REGIME IS MARKED IN BOLD

| Method        | 1% | Precision / Recall / F1 / IoU | 5% | Precision / Recall / F1 / IoU | 20% | Precision / Recall / F1 / IoU | 100% | Precision / Recall / F1 / IoU |
|---------------|----|-----------------------------|----|-----------------------------|-----|-----------------------------|------|-----------------------------|
|               | 73.54 | 34.58 / 47.05 / 30.76 | 83.00 | 77.76 / 80.30 / 67.08 | 90.50 | 84.94 / 87.63 / 77.98 |
| Random        | 81.40 | 53.45 / 65.18 / 48.35 | 89.59 | 82.26 / 85.76 / 75.08 | 91.29 | 85.25 / 88.16 / 78.83 |
| ImageNet      | 86.25 | 61.73 / 71.96 / 65.20 | 90.91 | 81.02 / 85.68 / 74.95 | 91.22 | 85.11 / 88.06 / 78.67 |
| In-domain     | 92.46 | 74.41 / 57.73 / 40.58 | 88.97 | 76.46 / 82.24 / 69.84 | 91.47 | 84.78 / 88.00 / 78.57 |
| SimSiam [21]  | 74.78 | 71.20 / 74.95 / 52.42 | 88.78 | 80.45 / 84.41 / 70.03 | 91.59 | 85.06 / 88.21 / 78.90 |
| MoCo-v2 [56]  | 91.51 | 58.62 / 71.46 / 55.60 | 89.33 | 82.09 / 85.56 / 74.76 | 94.63 | 79.98 / 86.69 / 76.51 |
| DenseCL [38]  | 84.45 | 56.40 / 67.65 / 51.09 | 90.30 | 77.94 / 83.67 / 71.92 | 90.73 | 86.59 / 88.16 / 79.55 |
| CMC [29]      | 84.74 | 61.43 / 71.23 / 53.51 | 90.34 | 83.70 / 84.02 / 72.44 | 91.03 | 85.68 / 88.27 / 79.01 |
| SeCo [16]     | 48.38 | 42.52 / 56.09 / 38.98 | 88.35 | 72.16 / 79.44 / 65.90 | 89.53 | 85.05 / 87.25 / 77.39 |
| Ours          | 78.50 | 66.29 / 71.88 / 56.10 | 86.45 | 75.04 / 80.34 / 67.15 | 90.15 | 81.37 / 85.53 / 74.73 |
|               | 78.81 | 66.62 / 72.15 / 56.43 | 83.93 | 76.43 / 80.00 / 66.67 | 89.29 | 86.93 / 88.09 / 78.72 |
|               | 72.89 | 54.51 / 62.37 / 45.32 | 72.89 | 54.51 / 62.37 / 45.32 | 91.80 | 65.02 / 76.12 / 61.45 |
|               | 88.93 | 59.59 / 71.36 / 55.48 | 94.99 | 76.81 / 84.94 / 73.82 | 92.36 | 84.69 / 87.90 / 78.41 |
|               | 69.82 | 60.43 / 64.79 / 47.91 | 88.96 | 73.11 / 80.26 / 67.03 | 91.36 | 84.69 / 87.90 / 78.41 |
|               | 67.29 | 61.14 / 64.07 / 47.03 | 91.80 | 72.60 / 81.08 / 68.18 | 90.73 | 84.69 / 87.90 / 78.41 |
|               | 81.24 | 61.80 / 70.20 / 54.08 | 90.23 | 76.69 / 82.91 / 70.81 | 93.19 | 71.36 / 80.83 / 67.82 |
|               | 77.36 | 57.88 / 66.21 / 49.49 | 92.20 | 76.72 / 83.75 / 72.04 | 89.56 | 85.47 / 87.66 / 78.03 |
|               | 73.97 | 66.64 / 70.01 / 53.86 | 92.20 | 76.72 / 83.75 / 72.04 | 92.51 | 76.98 / 84.03 / 72.46 |
|               | 80.08 | 60.62 / 69.00 / 52.68 | 80.08 | 60.62 / 69.00 / 52.68 | 80.08 | 60.62 / 69.00 / 52.68 |
|               | 85.32 | 67.06 / 75.09 / 60.12 | 85.32 | 67.06 / 75.09 / 60.12 | 92.13 | 81.06 / 86.24 / 75.81 |

Effectiveness of Transferable Features
We can observe that in-domain supervision shows relatively poorer results on the WHU-CD/Guangzhou-CD datasets than on the LEVIR-CD dataset. It may be due to the subdomain diversity, i.e., different RS datasets contain land-covers of diverse appearances and different spatial resolutions. The traditional supervised pretraining may overfit patterns of a certain subdomain and lack transferability to another.

2) Comparison to Other SSL Pretraining: Table I also shows that our method consistently outperforms several SSL methods, including two popular image-level representation learning methods, one DenseCL method, and two recent SSL methods in the RS domain. We can observe that the normal DenseCL can even be inferior to the image-level counterpart (MoCo) in many data conditions. It may be due to the discrepancy between the commonly used ImageNet image and the uncurated RS image. Unlike the ImageNet data that mainly contains one single object in the center of the image, multiple similar foreground objects may exist at different spatial locations in an RS image. The intrinsic characteristic of RS data brings difficulty in computing pixel correspondence across views in DenseCL. The resulting incorrect correspondence may hinder learning meaningful representations. On the contrary, we utilize the exact pixel correspondence for representation learning by leveraging spatial correspondences of the pixels in the overlapped area between views.

The comparison with CMC indicates that we exploit a better approach to using the semantic mask as additional information in a SSL framework to improve feature representations. We argue that directly using the mask as an augmented view along with the image views may not be an optimal solution. In [29], the mask is encoded into one vector that losses spatial details. Semantic spatial relations have not been fully utilized in such a scenario. Instead, we leverage semantic masks to push away representations of pixels from different semantic regions to learn semantic-aware features that can distinguish...
the foreground or background pixels, thus benefiting the recognition of objects of interest in the downstream CD task.

Visualization comparisons of the CD predictions on the three datasets for different pretrained methods are also given in Fig. 4. We use different colors for a better view, where white for true positives, black for true negatives, red for false positives, and green for false negatives. We could observe that the CD model using our pretrained model achieves more accurate predictions than those of its counterparts.

3) Better Convergence: We further give validation curves on the downstream CD task to show the transferability of each pretrained model. Fig. 5 depicts the curve of validation accuracy (mean F1-score) during the training phase using 5% LEVIR-CD/WHU-CD/Guangzhou-CD training data. Different colors denote different pretraining methods for a better view. It can be observed that our SaDL outperforms its counterparts in terms of stability and effectiveness. The comparison results indicate that our method provides better model initialization, which accelerates convergence and improves fine-tuning performance on the downstream CD task.

4) Computational Complexity: We also report the computational complexity of our method and other compared SSL methods. For a fair comparison, we test all the methods on a computing server equipped with an NVIDIA RTX 3090 GPU. Table II lists the number of model parameters (Params.), floating-point operations per second (FLOPs), and GPU inference time of each method. We set the size of the image input to the model to $256 \times 256 \times 3$. The reported time is the average time consumption of each image over 100 forward iterations given a random input sample of a batch size of 64. The results show that our method requires relatively higher FLOPs (about 1.5 times) than counterparts that have two views. It is because
our SaDL needs to calculate three views each iteration. We can also observe that our method has fewer FLOPs than SeCo, but takes slightly more time. It may be due to the additional time consumption of our online view synthesis and masked sampling process. Nevertheless, our computational overhead is reasonable and acceptable.

D. Ablation Studies

We conduct ablation studies on the three key components of our proposed SaDL: 1) dense representation learning via masked-based point sampling (MS); 2) learning discriminative features via SD; and 3) learning consistent foreground features via BS. We start from the SimSiam [21] with an FPN-based encoder, which learns consistent features between two artificially augmented views of the same image, as our baseline. We incrementally add the three key components to the baseline to evaluate their respective gains to the performance of the model. All the above models are trained on our pretraining dataset for 200 epochs and then transferred to the downstream CD task. Downstream experiments are performed under the 5% data regime of the LEVIR-CD/WHU-CD/Guangzhou-CD datasets.

1) Ablation on Masked Sampling (MS): Table III shows that our dense representation learning via sampling multiple points for each semantic category brings consistent and significant performance improvements across the three datasets, compared to the baseline. It indicates that our MS introduces much more valuable information to representation learning than the traditional global average pooling-based counterpart that loses spatial details. Moreover, we make a comparison to the masked pooling-based approach [53], which pools features belonging to each semantic region into one semantic vector by leveraging the semantic mask. Simply constraining each semantic vector invariant across views brings small improvement, compared to the global pooling approach (baseline). We can observe that our sampling-based method is far superior to the masked pool one. The multiple pixel correspondences across views through masked sampling provide more diverse supervision of spatial accordance, thus benefitting representation learning.

2) Ablation on SD: To make full use of the per-pixel semantic information, we employ an additional loss term, namely semantic dissimilarity loss, to learn discriminative features that can distinguish pixels from different semantic regions. Table III shows that our SD induces considerable improvements in the F1-score of the CD model across the three datasets. The empirical results indicate that our SD can significantly improve the model transferability. It may attribute to the role of the semantic mask in our SSL framework. We include the per-pixel semantic supervision into our pretraining in a contrastive manner. Different from directly mapping the image to the label, we learn an embedding space where pixels from the foreground (i.e., buildings) and the background are pushed apart. The resulting model obtains discriminative features that could distinguish land covers of interest from backgrounds, thus benefitting the recognition of change of interest in the downstream CD task.

3) Ablation on BS: Apart from two views by artificial augmentations, we propose a third view via swapping the background of the current image from that of another one. This incurs an additional cross-view similarity loss, i.e., learning invariant foreground features between view 1 and view 3. Quantitative results in Table III demonstrate that our BS further brings consistent improvements in the performance of the pretrained CD model on the three CD datasets. It indicates that the learned foreground representations invariant to background changes benefit the downstream CD for the land covers of interest. Our BS can be viewed as an approach to synthesizing unconcerned changes in a scene. Such design may enforce the model to focus more on changes of interest related to the foreground, but not irrelevant changes within the background.
To further verify the effectiveness of the introduced third view, we also compare the natural augmentation, i.e., augmenting the current view with a registered image from another temporal. For a fair comparison, we apply the same loss for constraining consistent foreground representations between view 1 and view 3. Table IV demonstrates our synthetic view consistently outperforms the natural counterpart. While natural augmentation provides rich illumination changes, undesired real foreground changes could also occur, which may hinder representation learning.

### E. Parametric Analysis

1) Sampled Point Number: The number of sampled points \( N \) for each semantic category is an important hyperparameter. We test different \( N \in \{1, 4, 16, 64\} \) to analyze its effect on the performance of the downstream fine-tuning performance. Table V reports the F1-score/IoU of the pretrained CD model trained on a 5% data regime of the LEVIR-CD/WHU-CD/Guangzhou-CD datasets. We can observe an overall improvement in the F1 score of the CD model on the three datasets when increasing the sample number from 1 to 16. It is because multiple sampled points from different spatial locations introduce more spatial-diverse information for both learning cross-view invariant and semantically discriminative features, thus facilitating the dense change recognition in the downstream CD task. We can also observe a drop in F1-score when further increasing \( N \) from 16 to 64. It may be because the over-sampled points nearby the edge of the semantic regions could bring difficulty in training semantically discriminative features. Therefore, we set \( N \) to 16.

### Table V

**Effect of Sampled Point Number \( N \). The F1-Score and IoU on the LEVIR-CD/WHU-CD/Guangzhou-CD Test Set are Reported. The CD Model Is Trained Under the Data Regime of 5%**

| \( N \) | LEVIR-CD F1 | LEVIR-CD IoU | WHU-CD F1 | WHU-CD IoU | Guangzhou-CD F1 | Guangzhou-CD IoU |
|-------|-------------|--------------|----------|-----------|----------------|------------------|
| 1     | 74.00       | 58.73        | 78.49    | 64.59     | 63.51          | 46.53            |
| 4     | 75.98       | 61.26        | 80.23    | 66.98     | 66.07          | 49.33            |
| 16    | **79.44**   | **65.90**    | **80.34**| **67.15** | 65.89          | 49.13            |
| 64    | 76.12       | 61.45        | 76.34    | 61.73     | 64.84          | 47.97            |

2) Pretraining Epochs Number: The epochs number \( N_e \) of pretraining affects its fine-tuning performance on the downstream task. We train our SaDL using varying epochs numbers \( N_e \in \{5, 20, 100, 200, 400\} \). Table VI reports the F1-score/IoU of the pretrained CD model on the LEVIR-CD/
WHU-CD/Guangzhou-CD datasets under a 5% data regime. Note that $N_e = 0$ denotes the ImageNet pretraining. We can observe a significant improvement in the F1-score even with a very small number of pretraining epochs (e.g., 5), compared to the baseline ($N_e = 0$). The performance of the pretrained CD model further improves with longer training (e.g., from 5 to 200 epochs). We can also observe a negligible improvement or even some drops in the F1-score of the CD model on the downstream tasks when further increasing the $N_e$ from 200 to 400. A pretraining with too many training epochs may overfit the pretraining data and incur poorer model transferability. Therefore, we set $N_e$ to 200.

F. Feature Visualization

1) Encoder Visualization: We incorporate semantic supervision into a self-supervised framework to learn semantically meaningful per-pixel representations. To better inspect the semantic relation of the learned representations, we visualize

---

Fig. 7. CAMs of different pretrained models. We test images from our pretraining dataset and the downstream LEVIR-CD/WHU-CD/Guangzhou-CD datasets.
the self-similarity map of the output dense features from our encoder.

**Feature self-similarity map:** For one selected point on an image, we can obtain one self-similarity map by calculating feature correlations between the selected point and all the pixels in the image. Formally, given the per-pixel features \( X \in \mathbb{R}^{HW \times C} \) and one selected point feature \( \mathbf{x} = X[p] \in \mathbb{R}^C \), the self-similarity map \( SM \in \mathbb{R}^{HW} \) of the point \( p \) is given by their dot product followed by softmax on the channel of \( HW \), i.e., \( SM = \text{softmax}(X \cdot x) \).

Fig. 6 shows some examples of self-similarity maps of the selected foreground points (viewed in red) on images in our pretraining dataset. Red denotes higher similarity values and blue denotes lower values. From Fig. 6(c), we can observe that the selected foreground point shows strong correlations with pixels of the same semantic category (building). It indicates that our SaDL can obtain high-resolution high-level features that reveal semantic relations between pixels.

Fig. 6 also gives visualization results for other ablation versions of the proposed model, including our baseline [f], baseline + MS [e], and baseline + MS + SD [d]. We can observe that ours with MS can highlight pixels nearby the selected point and shows a relatively more accurate location response than the baseline. It implies the effectiveness of the proposed dense representation learning based on masked sampling over the traditional average pooling approach. We can also observe our ablation [d] highlights more accurate foreground regions than that of e). It indicates our introduced SD incurs discriminative representations that well distinguish the foreground and the background. Moreover, our full SaDL model [c] obtain more refined visualization results by further adding BS to learn consistent foreground features invariant to background changes.

2) **Backbone Visualization:** For a fair comparison, we only transfer backbone (i.e., ResNet-18) parameters of our SaDL on the downstream CD model, like other pretrained models. To interpret what concepts are learned by the networks, we show the backbone visualization of each pretrained model by utilizing network visualization techniques-class activation maps (CAM) [83]. Basically, a CAM is the weighted sum of each channel activation map from a certain layer in the network. Here, we visualize the last convolution layer of the backbone.

Fig. 7 shows the comparison of CAM for different pretrained models. Red denotes higher attention values and blue denotes lower values. We test images from our pretraining dataset as well as the three downstream CD datasets. We can observe that the CAM of our model shows high concentrations at the foreground objects in the image with a satisfactory localization accuracy. It is notable that our model can not only attend to land covers of interest on the pretraining images but also perform well on the downstream images with considerable domain gaps (e.g., differences in spatial resolution and object style) from the pretraining data. For instance, the high-density buildings (see the last row in Fig. 7) in the Guangzhou-CD dataset are hardly present in the pretraining dataset. It indicates that our proposed pretraining could extract transferable semantically meaningful representations with well generalization ability.

**V. Discussion**

We present semantic-aware representation learning for CD in high-resolution RS images. One major challenge of CD is how to improve the model generalization ability under the insufficient training labeled data regime. The issue of label insufficiency comes from two aspects: 1) imaging and land-cover diversity making it impossible to collect training samples covering real-world conditions and 2) extensive labor to collect bitemporal images containing changes and to annotate per-pixel labels. We leverage existing freely available data with semantic information in the RS community to learn a pretrained model to benefit the downstream CD task. Considering the key to CD is to recognize the real changes and exclude other irrelevant changes, we incorporate such CD task priors in the pretraining phase by learning semantically meaningful representations that could distinguish foreground land-covers from other backgrounds, and that are invariant to irrelevant changes, e.g., illumination differences and unconcerned land-cover changes. We explore semantic-supervised pretraining in a contrastive manner, which is different from traditional supervised pretraining that learns the mapping from the image to the label or other SSL methods that lack of semantic supervision. Compared to several existing pretrained methods, our method significantly improves the CD data efficiency (see Fig. 4) and could extract semantically meaningful representations with good generalization ability even on the unseen domain (see Fig. 7). It indicates our pretrained method can learn transferable discriminative features that benefit change recognition and incur better model generalization in the downstream CD task.

**VI. Conclusion**

In this article, we proposed a semantic-aware pretraining method for RS image CD. We incorporate the semantic information into a SSL framework to enhance the feature discrimination ability. Instead of manipulating the image-level features, we constrain pixel-level representation consistency based on sampling multiple spatially aligned points across views by introducing location correspondences. We employ an additional loss term to learn discriminative features that can distinguish foreground land covers and other backgrounds. Apart from two views via artificial augmentations, we generate a third view via BS to learn consistent presentations invariant to irrelevant changes. Extensive experiments on three downstream CD datasets verify the effectiveness of the proposed method. Our SaDL significantly outperforms ImageNet pretraining, in-domain supervised pretraining, and several SSL pretraining methods, especially in small data regimes. The empirical results indicate that SaDL can well alleviate the labeled data insufficiency in CD. Notably, we can achieve comparable or even better results using only 20% training data than the random initiation baseline using 100% data. The feature visualization results also demonstrate that our method can extract semantically meaningful representations with well
generalization ability on different downstream images with domain discrepancies from the pretraining data.

Modern machine learning-based models typically suffer from spectral variabilities [84] in RS data, caused by different imaging conditions, including illuminations, atmospheric effects, and instrumental configurations. Our proposed pre-trained model could give consistent representations invariant to irrelevant changes (e.g., illumination and uncorrelated land cover changes), to benefit the downstream CD task. In the future, we will explore effective ways to explicitly model more kinds of imaging conditions in our pretraining framework to better handle the spectral variabilities. Other future directions include: 1) sampling strategies. Instead of treating each pixel equally, we could leverage the guidance map generated by edge clues or network attention maps to sample important pixels that may better benefit the representation learning; 2) exploring various pretraining backbones, e.g., more kinds of CNN structures and even other vision transformer architectures; and 3) extending our semantic-aware pretraining to distinguish more land cover types, which may benefit the multiclass semantic CD task.

REFERENCES

[1] A. Singh, “Review article digital change detection techniques using remotely-sensed data,” Int. J. Remote Sens., vol. 10, no. 6, pp. 989–1003, 1989, doi: 10.1080/014311689093939.

[2] 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. 6, pp. 1662, 2020.

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

[4] P. de Bern, O. de Carvalho Junior, R. Fontes Guimarães, and R. Trancoso Gomes, “Change detection of deforestation in the Brazilian Amazon using landart data and convolutional neural networks,” Remote Sens., vol. 12, no. 6, p. 901, Mar. 2020.

[5] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, “Building damage detection in satellite imagery using convolutional neural networks,” 2019, arXiv:1910.06444.

[6] L. Bruzzone and F. Bovolo, “A novel framework for the design of change detection algorithms for very-high-resolution remote sensing images,” Proc. IEEE, vol. 101, no. 3, pp. 609–630, Mar. 2013.

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

[8] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, “Triplet-based semantic relation learning for aerial remote sensing image change detection,” IEEE Geosci. Remote Sens. Lett., vol. 16, no. 2, pp. 266–270, Feb. 2019.

[9] F. I. Diakogiannis, F. Waldner, and P. Caccetta, “Looking for change? Roll the dice and demand attention,” Remote Sens., vol. 13, no. 18, p. 3707, 2021.

[10] 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.

[11] J. Chen et al., “DANet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 1194–1206, 2021.

[12] S. Saha, F. Bovolo, and L. Bruzzone, “Building change detection in VHR SAR images via unsupervised deep transcooding,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 3, pp. 1917–1929, Mar. 2021.

[13] F. Jiang, M. Gong, T. Zhan, and X. Fan, “A semisupervised GAN-based multiple change detection framework in multi-spectral images,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 7, pp. 1223–1227, Jul. 2020.

[14] J. Liu et al., “Convolutional neural network-based transfer learning for optical aerial images change detection,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 1, pp. 127–131, Jan. 2020.

[15] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, “An empirical study of remote sensing pretraining,” 2022, arXiv:2204.02825.
Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning in remote sensing images via local–global pyramid network and cross-task transfer learning strategy,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4704817.

T. Lei et al., “Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

S. Hafner, Y. Ban, and A. Nascetti, “Urban change detection using a dual-task Siamese network and semi-supervised learning,” 2022, arXiv:2204.12202.

W. Gedara Chaminda Bandara and V. M. Patel, “Revisiting consistency regularization for semi-supervised change detection in remote sensing images,” 2022, arXiv:2204.08544.

V. Ruzicka, S. D’Aronco, J. D. Wegner, and K. Schindler, “Deep active learning in remote sensing for data efficient change detection,” in *Proc. MACLEAN: Mach. Learn. Earth Obs. Workshop Co-located Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Databases (ECML/PKDD)* (CEUR Workshop Proceedings), vol. 2766, T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, Eds., 2020, pp. 1–16, 2022.

T. Chen, S. Deshmukh, and A. Stoian, “Active learning for interactive satellite image change detection,” 2021, arXiv:2110.04250.

V. Risoevič and V. Stojnić, “Do we still need ImageNet pre-training in remote sensing scene classification?” 2021, arXiv:2111.03690.

M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, “Training general representations for remote sensing using in-domain knowledge,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.* (IGARSS), Honolulu, HI, USA: AAAI Press, Jan./Feb. 2019, pp. 3967–3974.

H. Sahbi, S. Deshmukh, and A. Stoian, “Active learning for interactive satellite image change detection,” 2022, arXiv:2201.05778.

X. Liu et al., “Self-supervised learning: Generative or contrastive,” *IEEE Trans. Knowl. Data Eng.*, early access, Jun. 22, 2021, doi: 10.1109/TKDE.2021.3098866.

T. Chen, Sonnbühlt, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. 37th Int. Conf. Mach. Learn. (ICML) (Proceedings of Machine Learning Research)*, vol. 119. JMLR.org, 2020, pp. 1507–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20h.html.

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

J. Grill et al., “Boost your own latent—A new approach to self-supervised learning,” in *Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Red Hook, NY, USA: Curran Associates, Dec. 2020, pp. 1–14. [Online]. Available: http://proceedings.neurips.cc/paper/2020/hash/1fa3aad8054c4ec70142b178b1929259e-Abstract.html.

X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, “Dense contrastive learning for self-supervised visual pre-training,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), Jun. 2021, pp. 3023–3032.

Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jun. 2018, pp. 3733–3742, doi: 10.1109/CVPR.2018.00393.

M. Caron et al., “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

N. Jean, S. Wang, A. Samar, G. Azzari, D. B. Lobell, and S. Ermon, “Tile2Vec: Unsupervised representation learning for spatially distributed data,” in *Proc. 33rd AAAI Conf. Artif. Intell.*, AAAI, 31st Innov. Appl. Intell. Conf., IAAI, 9th AAAI Symp. Educ. Adv. Intell. (EAAI), Honolulu, HI, USA: AAAI Press, Jan./Feb. 2019, pp. 3967–3974.

C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, “Remote sensing image scene classification with self-supervised paradigm under limited labeled samples,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 8004005.

B. Peng, Q. Huang, and J. Rao, “Spatiotemporal contrastive representation learning for building damage classification,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.* (IGARSS), Brussels, Belgium: IEEE, Jul. 2021, pp. 8562–8565.

P. Jain, B. Schoen-Phelan, and R.ross, “Self-supervised learning for invariant representations from multi-spectral and SAR images,” 2022, arXiv:2205.02049.

Y. Chen and L. Bruzzone, “Self-supervised change detection in multiview remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

S. Saha, P. Ebel, and X. X. Zhu, “Self-supervised multisensor change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.

Y. Chen and L. Bruzzone, “Self-supervised remote sensing images change detection at pixel-level,” 2021, arXiv:2105.08501.

Y. Zhang, G. Tao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, “Graph convolutional networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

D. Hong et al., “SpectralFormer: Rethinking hyperspectral image classification with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

X. X. Zhu et al., “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

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

X. Wu, D. Hong, and J. Chanussot, “Convolutional neural networks for multimodal remote sensing data classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.

J. Kang, S. Tariq, H. Oh, and S. S. Woo, “A survey of deep learning-based object detection methods and datasets for overhead imagery,” *IEEE Access*, vol. 10, pp. 20118–20134, 2022.

J. Wu et al., “A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102615. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243421003226.

J. Qu, Y. Xu, W. Dong, Y. Li, and Q. Du, “Dual-branch difference amplification graph convolutional network for hyperspectral image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

W. Gedara Chaminda Bandara and V. M. Patel, “A transformer-based Siamese network for change detection,” 2022, arXiv:2201.01293.

X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1501–1510.

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

T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

Y. Tian, D. Krishman, and P. Isola, “Contrastive multiview coding,” in *Proc. 16th Eur. Conf. Comput. Vis.* Berlin, Germany: Springer-Verlag, Aug. 2020, pp. 776–794, doi: 10.1007/978-3-030-58621-8_45.
H. Wang et al., “Score-CAM: Score-weighted visual explanations for convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Los Alamitos, CA, USA, Jun. 2020, pp. 111–119, doi: 10.1109/CVPRW50498.2020.00020.

D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “An augmented linear mixing model to address spectral variability for hyperspectral unmixing,” IEEE Trans. Image Process., vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

Hao Chen (Graduate Student Member, IEEE) received the B.S. degree from the Image Processing Center, School of Astronautics, Beihang University, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree.

His research interests include machine learning, deep learning, and semantic segmentation.

Wenyuan Li received the B.S. degree from North China Electric Power University, Beijing, China, in 2017. He is currently pursuing the Ph.D. degree with the Image Processing Center, School of Astronautics, Beihang University, Beijing.

His research interests include deep learning, image processing, and pattern recognition.

Song Chen received the master’s degree from the School of Journalism and Communication, Jeonbuk National University, Jeonju-si, South Korea, in 2018, where he is currently pursuing the Ph.D. degree in journalism and communication.

His research interests include new media language, social media, and web media.

Zhenwei Shi (Member, IEEE) received the Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, from 2013 to 2014. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Associate Editor for the Infrared Physics and Technology. He has authored or coauthored over 100 scientific articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the IEEE Conference on Computer Vision and Pattern Recognition. His personal website is http://levir.buaa.edu.cn/.