

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

Hao Chen, 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 pre-trained models is effective to alleviate label insufficiency in remote sensing (RS) change detection (CD). We explore the use of semantic information during pre-training. Different from traditional supervised pre-training 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 pre-training. Extensive experiments on three CD datasets verify the effectiveness of our method. Ours significantly outperforms ImageNet pre-training, in-domain supervision, and several SSL methods. Empirical results indicate our pre-training 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—Remote sensing image, change detection (CD), self-supervised learning (SSL), semantic-aware representation learning, convolutional neural networks.

I. INTRODUCTION

REMOTE sensing (RS) image change detection (CD) is the process of identifying changes of interest in a scene based on multi-temporal 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.

This work was supported by the National Natural Science Foundation of China under Grant 62125102. (Corresponding author: Zhenwei Shi)

Hao Chen (email: justchenhao@buaa.edu.cn), Wenyuan Li (email: liwenyuan@buaa.edu.cn) and Zhenwei Shi (Corresponding Author, e-mail: shizhenwei@buaa.edu.cn) are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China. Song Chen (email: justharrychen@naver.com) is with Department of Journalism and Communications, Jeonbuk National University, Jeonju-si 54896, South Korea.

The available multi-temporal 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, 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 pre-trained model previously trained on a large-scale dataset onto a downstream task that contains relatively small labeled data. ImageNet pre-training is commonly applied in contemporary CD models [8–13]. Considering the domain gap between natural and RS images may induce sub-optimal representations, a new trend is to pre-train on RS data to learn in-domain representations, including supervised RS pre-training [10, 14, 15] and unsupervised RS pre-training based on self-supervised learning (SSL) [16–19]. In this work, we focus on supervised RS pre-training via leveraging existing RS data in the RS community. Different from existing supervised pre-training 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 [16–18, 22–33] and have shown promising performance for the downstream supervised CD task [16–18].

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 remote sensing 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. Towards 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 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 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 pre-training 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 change detection datasets [2, 35, 36]. Extensive experiments on the downstream CD task verify the effectiveness of the proposed pre-training.

The contribution of our work can be summarised as follows:

- Different from traditional supervised pre-training 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 pre-training 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 semantic discrimination to learn spatially-sensitive features, thus benefiting the downstream dense prediction CD task.
- 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 background swap version to learn representations invariant to irrelevant changes (unconcerned land covers and illumination).
- Extensive experiments on three CD datasets verify the effectiveness of the proposed method. Ours outperforms

ImageNet pre-training, in-domain supervision, and several state-of-the-art SSL methods.

The rest of this paper is organized as follows. Sec. II introduces related work of CD data-efficiency approaches and pre-training methods. Sec. III describes our proposed semantic-aware pre-training method. Experimental results are given in Sec. IV, and the Conclusion is drawn in Sec. 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[37–43], generating pseudo labels for unlabeled data via semi-supervised learning [36, 44, 45], using active learning to select a small number of informative samples [46–49], and transfer learning from pre-trained 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, 39–43], including geometric transformations (e.g., random crop, horizontal flip), color transformations, and Gaussian blur, etc. A recent advance increases the number of the positive samples (change of interest) by blending the GAN-generated instance on the appropriate spatial-temporal position of the bitemporal image [38].

Transfer learning from a pre-trained model allows leveraging the knowledge of a source dataset to improve the data efficiency of a downstream task. ImageNet pre-training is widely used in change detection [7–13] 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 pre-train on the remote sensing data to learn the in-domain representations [10, 14–17, 19, 50–53]. Supervised RS pre-training normally learns discriminative representations from the RS image-level [10, 15] or pixel-level [14] land-cover classification samples, while self-supervised pre-training could learn meaningful representations by making use of a vast amount of unlabeled RS data [16, 17, 19].

Our method falls in the supervised pre-training scope. Different from normal supervised pre-training 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 pre-training.

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

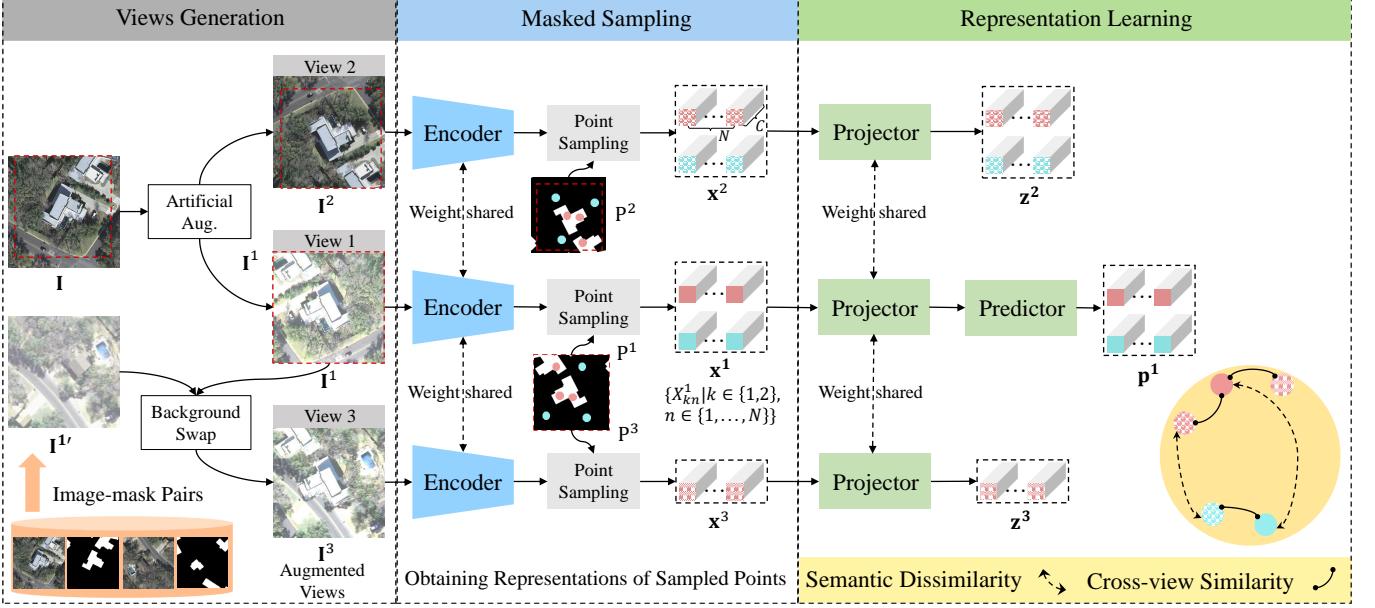


Fig. 1. Illustration of our semantic-aware dense representation learning. Our framework follows SimSiam [21] to use simple Siamese networks to learn meaningful image representations. Instead of constraining image-level consistency, we achieve point-level cross-view consistency to sustain spatially sensitive features. An equal number of points for each category are sampled from the overlapped region between views by leveraging the semantic mask. We learn an embedding space where points from different semantic regions are pushed apart and those spatially aligned points across views are pulled together. Apart from two views via basic artificial augmentations, we generate a third view via the background swap, thus learning illumination invariant representations as well as consistent foreground representations invariant to uninterested background changes. For simplicity, we omit the symmetrized cross-view similarity loss computed by exchanging the two views.

B. Contrastive pre-training methods

Self-supervised learning 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., SimCLR [55], MoCo [20, 56], BYOL [57], SimSiam [21] and DenseCL [58]) have made significant progress in SSL. The key to contrastive approaches is to perform the instance discrimination pre-text 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 pre-training have been shown to benefit RS downstream tasks, including land-cover classification [22, 24–28, 61, 62], semantic segmentation [29–32], and change detection [16–18]. Most existing methods apply InfoNCE loss [18, 22, 24–28, 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., multi-temporal co-registered images) [16, 17, 22, 26, 63], and modality augmentations (e.g., optical image, 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 change detection have seldom been explored. Some attempts apply SSL directly on a small downstream change detection dataset to extract seasonal invariant features for unsupervised change detection [65–67]. Differently, we employ an additional relatively large-scale RS dataset to learn semantic-aware representations and then transfer knowledge from the pre-trained 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 pre-trained model on the medium-resolution change detection dataset [16–18]. We instead explore pre-trained models suitable for high-resolution RS image change detection. Instead of manipulating global representations, we propose dense points consistency based on spatial correspondence between the positive pairs 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. 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, 65–67] treat co-registered multi-temporal images as positive samples to learn the seasonal invariant features. However, the intrinsic change of interest within the bitemporal image may induce the pre-trained model invariant to real changes, which may be unfavorable to the downstream CD task. Different from existing methods using natural bitemporal images, ours learns 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. Deep learning-based models in remote sensing

The recent few years witness the great success of deep learning 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/semantic segmentation [68, 69], multimodal image classification [71, 72], object detection [73], and change detection [2, 7, 37–43, 74–77]. CNNs have been demonstrated to be effective to extract meaningful and transferable image representations by multiple levels of abstraction, thus attracting much attention in remote sensing society. Most of the existing deep learning-based CD models [2, 37–43] 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 change detection [74–77], for improving model discriminative ability by leveraging their capacity of modeling long-range relations among image pixels. In this paper, 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 pre-training 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 Sec. IV-A) in a painless manner for contrastive pre-training. Specially, we focus on the building category as the objects of interest to perform change detection.

We propose Semantic-aware Dense representation Learning (SaDL) for remote sensing image change detection. 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, 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 self-supervised learning 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 Sec. 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 Sec. III-C.

3) **Representation Learning.** Given the sampled point representations, we employ 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 Sec. III-D.

B. Views Generation

Given each input training sample (including an image \mathbf{I} and a mask \mathbf{M}) in a mini-batch, we produce three views of the same image, including two basic transformed versions, and one background swap version. Let $\mathcal{T} = \mathcal{T}_g \circ \mathcal{T}_c$ denotes a set of artificial augmentations, including color augmentations \mathcal{T}_c (e.g., color jittering and Gaussian blur) and geometric augmentations \mathcal{T}_g (e.g., random cropping and flipping). Details of our views generations are described in Algorithm 1.

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

$$\begin{aligned}\mathbf{I}^1 &= \mathcal{T}(\mathbf{I}) = \mathcal{T}_g(\mathcal{T}_c(\mathbf{I})) \\ \mathbf{I}^2 &= \mathcal{T}'(\mathbf{I}) = \mathcal{T}'_g(\mathcal{T}'_c(\mathbf{I}))\end{aligned}\quad (1)$$

where $\mathcal{T}, \mathcal{T}'$ denote two random sets of artificial augmentations.

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

$$\begin{aligned}\mathbf{M}^1 &= \mathcal{T}_g(\mathbf{M}) \\ \mathbf{M}^2 &= \mathcal{T}'_g(\mathbf{M}).\end{aligned}\quad (2)$$

View generation via background swap. The third view of \mathbf{I} is obtained by swapping the background of the first view \mathbf{I}^1 with that of another image \mathbf{I}' . For simplicity, \mathbf{I}' has the

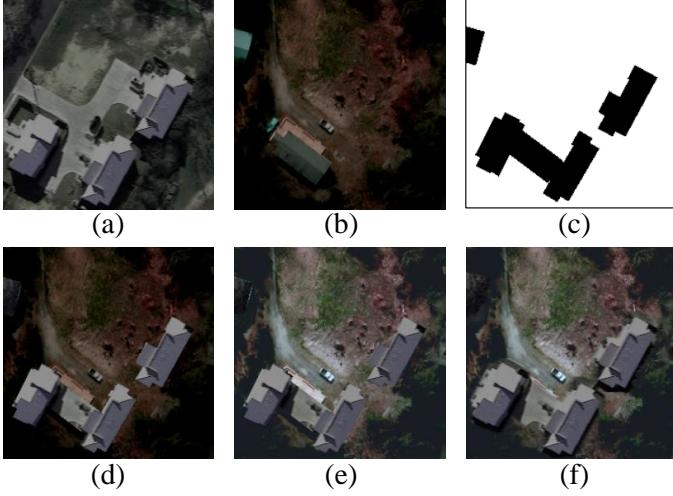


Fig. 2. Illustration of our view synthesis. View 3 is generated by swapping the background of (a) the view 1 of the current image $I^{1'}$ and that of (b) another image $I^{1'}$ using (c) their common background mask M_{bg} . We compare different compositions: (d) direct pasting, (e) with color transfer, (f) with color transfer + context-aware blending.

inverted index as \mathbf{I} from the same mini-batch. In other words, the summation of the index of \mathbf{I} and that of \mathbf{I}' equals the mini-batch size. Note that we swap the common background area between $\mathbf{I}^{1'}$ and \mathbf{I}^1 to simulate the changes of unconcerned land covers. We obtain the common background mask M_{bg} by calculating the intersection of the background regions in \mathbf{M}^1 and $\mathbf{M}^{1'}$.

However, directly pasting the background of $\mathbf{I}^{1'}$ on \mathbf{I}^1 using M_{bg} may result in undesired synthetic artifacts (see Fig. 2 (d)) that could be detrimental to representation learning. Towards 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 $\mathbf{I}^{1'}$ (source image) and \mathbf{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

$$\mathbf{I}_t^{1'} = \sigma(\mathbf{I}^1) \left(\frac{\mathbf{I}^{1'} - \mu(\mathbf{I}^{1'})}{\sigma(\mathbf{I}^{1'})} \right) + \mu(\mathbf{I}^1) \quad (3)$$

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 \mathbf{I}^1 by performing morphological erosion on the background mask. We further blur the eroded background mask using a Gaussian filter towards natural and seamless composition. The composite image \mathbf{I}^3 (view 3) is calculated by alpha blending using the eroded mask of the blurred version:

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

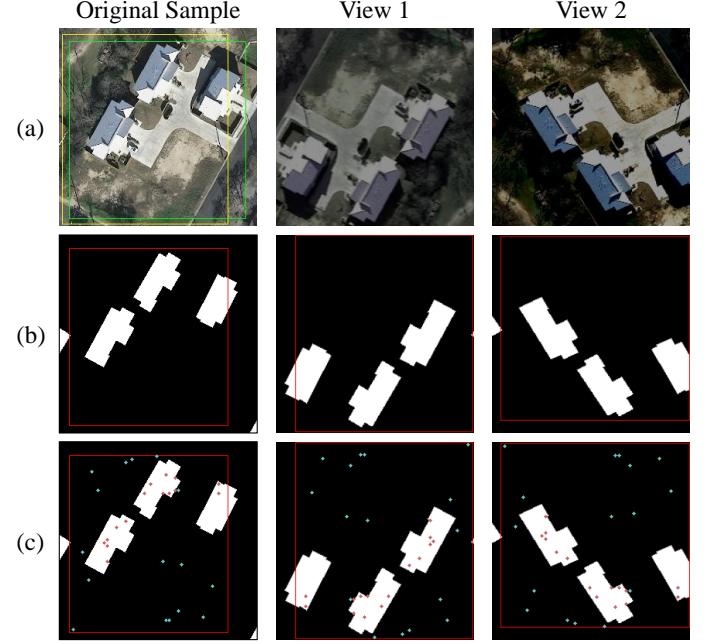


Fig. 3. Illustration of our masked sampling. (a) Images. View 1/2 corresponds to the region within the green/yellow bounding box in the original image. (b) Semantic masks. The overlapped region between views is shown in red. (c) Sampled points. The same number of points for each semantic region in the overlapped area are randomly sampled. Different colors are used for each category for a better view.

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

Obtaining dense representations. The generated three views $\mathbf{I}^1, \mathbf{I}^2, \mathbf{I}^3$ are then encoded by a Siamese neural network f into dense representations $\mathbf{X}^i = f(\mathbf{I}^i) \in \mathbb{R}^{H \times W \times C}, i \in \{1, 2, 3\}$. 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 Sec. 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 features may better reduce the gap between the pre-training 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.

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. In order to perform cross-view point consistency, we first need to obtain the overlapped region between \mathbf{I}^1 and \mathbf{I}^2 . Given the known geometry augmentations (\mathcal{T}_g

Algorithm 1: Views generation for a batch of samples

Input: $\hat{\mathbf{I}} \in \mathbb{R}^{B \times H_0 \times W_0 \times 3}$ (a batch of input images)
Input: $\hat{\mathbf{M}} \in \mathbb{R}^{B \times H_0 \times W_0}$ (a batch of corresponding semantic masks)
Output: $\hat{\mathbf{I}}^1, \hat{\mathbf{I}}^2, \hat{\mathbf{I}}^3 \in \mathbb{R}^{B \times H_0 \times W_0 \times 3}$ (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    $\hat{\mathbf{I}}^1[b], \hat{\mathbf{I}}^2[b] \leftarrow \mathcal{T}(\hat{\mathbf{I}}[b]), \mathcal{T}'(\hat{\mathbf{I}}[b])$ 
5    $\hat{\mathbf{M}}^1[b], \hat{\mathbf{M}}^2[b] \leftarrow \mathcal{T}_g(\hat{\mathbf{M}}[b]), \mathcal{T}'_g(\hat{\mathbf{M}}[b])$ 
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   $\mathbf{I}' \leftarrow \hat{\mathbf{I}}^1[B - b]$ 
11   $\mathbf{M}' \leftarrow \hat{\mathbf{M}}^1[B - b]$ 
12   $\mathbf{I}'_t \leftarrow \text{ColorTransfer}(\mathbf{I}', \hat{\mathbf{I}}^1[b])$ 
13  // obtain the common background mask
14   $\mathbf{M}_{bg} \leftarrow (\hat{\mathbf{M}}^1[b] == 0) \cdot (\mathbf{M}' == 0)$ 
15  // context-aware blending
16   $\mathbf{M}_{bg} \leftarrow \text{Erosion}(\mathbf{M}_{bg})$ 
17   $\hat{\mathbf{I}}^3[b] \leftarrow (1 - \mathbf{M}_{bg})\hat{\mathbf{I}}^1 + \mathbf{M}_{bg}\mathbf{I}'_t$ 
18 end

```

and \mathcal{T}'_g), we could calculate the corresponding bounding box region ($BB^i = (u^i, v^i, w^i, h^i)$) of view $i, i \in \{1, 2\}$ in the original image \mathbf{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 (\mathbf{I}^1) and view 2 (\mathbf{I}^2) respectively correspond to the region within the green/yellow bounding box in the original image \mathbf{I} . The overlapped area between views in form of bounding boxes (red) is shown in Fig. 3 (b).

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 (\mathbf{I}/\mathbf{M}). N is the number of sampled points for each class. Our default N is set to 16 (See Sec. III-E). $k=1$ denotes the background and $k=2$ denotes the foreground class. Let $P = \{P_{kn} = (u_{kn}, v_{kn}) | n \in \{1, 2, \dots, N\}, k \in \{1, 2\}\}$ denotes the sampled point set. The corresponding points in view 1 and view 2 can be obtained by coordinate transformation

Algorithm 2: Dense representation learning based on masked sampling for one sample

Input: $\mathbf{I}^1, \mathbf{I}^2, \mathbf{I}^3 \in \mathbb{R}^{B \times H_0 \times W_0 \times 3}$ (image view 1/2/3)
Input: $\mathbf{M} \in \mathbb{R}^{B \times H_0 \times W_0}$ (semantic mask of the image)
Input: $\mathcal{T}_g, \mathcal{T}'_g$ (applied geometry augmentations for the input)
Output: $\mathcal{L} \in \mathbb{R}$ (the loss on the input sample)

```

1 // reverse the area in input corresponding to view 1/2
2  $BB^1, BB^2 \leftarrow \text{Reverse}(\mathcal{T}_g), \text{Reverse}(\mathcal{T}'_g)$ 
3 // obtain overlapped region (BB) between views
4  $BB \leftarrow \text{Overlap}(BB^1, BB^2)$ 
5 // sampling N points in BB for each semantic category
6 for k in 1 : 2 do
7   for n in 1 : N do
8     sample  $P_{kn} \sim \text{Uniform}(\{p = (u, v) | p \text{ in } BB, \mathbf{M}[u][v] = k - 1\})$ 
9   end
10 end
11 // define the sampled point set
12  $P \triangleq \{P_{kn} | n \in \{1, 2, \dots, N\}, k \in \{1, 2\}\}$ 
13 // obtain cooresponding points in each view
14  $P^1, P^2 \leftarrow \mathcal{T}_g(P), \mathcal{T}'_g(P)$ 
15  $P^3 \leftarrow \{P_{kn}^1 | n \in \{1, 2, \dots, N\}, k = 2\}$ 
16 // coordinates downscaling
17  $P^1, P^2, P^3 \leftarrow P^1 // ds, P^2 // ds, P^3 // ds$ 
18 //obtain point representations for each view
19 for i in 1 : 3 do
20    $\mathbf{X}^i \leftarrow f(\mathbf{I}^i)$ 
21    $\mathbf{x}^i \leftarrow \text{gather}(\mathbf{X}^i, P^i)$ 
22   // obtain projections/predictions
23    $\mathbf{z}^i \leftarrow g(\mathbf{x}^i)$ 
24    $\mathbf{p}^i \leftarrow h(\mathbf{z}^i)$ 
25 end
26 // compute losses
27 // loss1: distinguish foreground and background
28 for i in 1 : 2 do
29   for n in 1 : N do
30      $\mathcal{L}_{sd,n}^i \leftarrow \mathcal{D}(\mathbf{x}_{1n}^i, \mathbf{x}_{2n}^i) + 1$ 
31   end
32 end
33  $\mathcal{L}_{sd} \leftarrow \frac{1}{2N} \sum_{i=1}^2 \sum_{n=1}^N \mathcal{L}_{sd,n}^i$ 
34 // loss2: consistency between view 1/2
35 for k in 1 : 2 do
36   for n in 1 : N do
37      $\mathcal{L}_{s,kn}^{(1,2)} \leftarrow 1 - \frac{1}{2}(\mathcal{D}(\mathbf{p}_{kn}^1, \text{stopgrad}(\mathbf{z}_{kn}^2)) +$ 
38      $\mathcal{D}(\mathbf{p}_{kn}^2, \text{stopgrad}(\mathbf{z}_{kn}^1)))$ 
39   end
40 end
41  $\mathcal{L}_{s1} \leftarrow \frac{1}{2N} \sum_{k=1}^2 \sum_{n=1}^N \mathcal{L}_{s,kn}^{(1,2)}$ 
42 // loss3: foreground consistency between view 1/3
43 for n in 1 : N do
44    $\mathcal{L}_{s,2n}^{(1,3)} \leftarrow 1 - \frac{1}{2}(\mathcal{D}(\mathbf{p}_{2n}^1, \text{stopgrad}(\mathbf{z}_{2n}^3)) +$ 
45    $\mathcal{D}(\mathbf{p}_{2n}^3, \text{stopgrad}(\mathbf{z}_{2n}^1)))$ 
46 end
47  $\mathcal{L}_{s2} \leftarrow \frac{1}{N} \sum_{n=1}^N \mathcal{L}_{s,2n}^{(1,3)}$ 
48  $\mathcal{L} \leftarrow \mathcal{L}_{sd} + \mathcal{L}_{s1} + \mathcal{L}_{s2}$ 

```

under the given geometry augmentations:

$$\begin{aligned} P^1 &= \mathcal{T}_g(P) \\ P^2 &= \mathcal{T}'_g(P) \end{aligned} \quad (5)$$

where $P^i = \{P_{kn}^i = (u_{kn}^i, v_{kn}^i) | n \in \{1, 2, \dots, N\}, k \in \{1, 2\}\}$ denotes the set of coordinates of the sampled point in view $i, i \in \{1, 2\}$. Note that we divide P^i by d_s (i.e., the down-sampling 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 $P^3 = \{P_{kn}^3 = P_{kn}^1 | n \in \{1, 2, \dots, N\}, k = 2\}$.

Then for each view $i, i \in \{1, 2\}$, we gather N representations \mathbf{x}_{kn}^i in the embedding space $\mathcal{V} \in \mathbb{R}^C$ for each semantic category k by indexing points on the feature map \mathbf{x}^i using coordinates P^i :

$$\mathbf{x}^i = \text{gather}(\mathbf{X}^i, P^i) \quad (6)$$

where $\mathbf{x}^i = \{\mathbf{x}_{kn}^i | n \in \{1, 2, \dots, N\}, k \in \{1, 2\}\}$ represents the set of representations of sampled points for view i .

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

D. Representation Learning

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 \mathcal{V} . For simplicity, we construct N pairs of points by one-to-one correspondence between N foreground points \mathbf{x}_{2n}^i and N background points $\mathbf{x}_{1n}^i, n \in \{1, \dots, N\}$ for each view $i \in \{1, 2\}$. Note that we supplement 1 to cosine similarity to ensure a non-negative loss value. Our semantic dissimilar loss $\mathcal{L}_{sd,n}^i$ on one constructed point pair indexed n for view i is defined as follows:

$$\mathcal{L}_{sd,n}^i = \mathcal{D}(\mathbf{x}_{1n}^i, \mathbf{x}_{2n}^i) + 1 \quad (7)$$

where $\mathcal{D}(\cdot, \cdot)$ denotes the cosine similarity, given by:

$$\mathcal{D}(a, b) = \frac{a}{\|a\|_2} \cdot \frac{b}{\|b\|_2} \quad (8)$$

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

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

$$\mathcal{L}_{sd} = \frac{1}{2N} \sum_{i=1}^2 \sum_{n=1}^N \mathcal{L}_{sd,n}^i. \quad (9)$$

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 \mathbf{x}_{kn}^i from the encoder are projected into a subspace $\mathcal{Z} \in \mathbb{R}^{C'}$ by a projection MLP head (projector) $g : \mathcal{V} \mapsto \mathcal{Z}$ to obtain projection vectors $\mathbf{z}_{kn}^i = g(\mathbf{x}_{kn}^i)$. Then, a prediction MLP head (predictor) $h : \mathcal{Z} \mapsto \mathcal{Z}$ transforms the output of one view (i.e., \mathbf{z}_{kn}^i) to obtain a prediction vector (i.e., $\mathbf{p}_{kn}^i = h(\mathbf{z}_{kn}^i) \in \mathbb{R}^{C'}$) in the same space and matches it to the other view (i.e., $\mathbf{z}_{kn}^{i'}$). 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:

$$\begin{aligned} \mathcal{L}_{s, kn}^{(i, i')} &= 1 - \frac{1}{2} (\mathcal{D}(\mathbf{p}_{kn}^i, \text{stopgrad}(\mathbf{z}_{kn}^{i'})) + \\ &\quad \mathcal{D}(\mathbf{p}_{kn}^{i'}, \text{stopgrad}(\mathbf{z}_{kn}^i))) \end{aligned} \quad (10)$$

where $\text{stopgrad}(\cdot)$ denotes the stop-gradient operation. The encoder and projector can not 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 $n \in \{1, \dots, N\}$ for each semantic category $k \in \{1, 2\}$:

$$\mathcal{L}_{s1} = \frac{1}{2N} \sum_{k=1}^2 \sum_{n=1}^N \mathcal{L}_{s, kn}^{(1, 2)}. \quad (11)$$

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

$$\mathcal{L}_{s2} = \frac{1}{N} \sum_{n=1}^N \mathcal{L}_{s, 2n}^{(1, 3)}. \quad (12)$$

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:

$$\mathcal{L} = \mathcal{L}_{sd} + \mathcal{L}_{s1} + \mathcal{L}_{s2}. \quad (13)$$

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

E. Implementation Details

Encoder. We use the ResNet-18 [79] backbone (without fully connected layers and global pooling) equipped with Feature Pyramid Networks (FPN) [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 1×1 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.

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

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.

Data augmentation. We set probability of color jittering to 0.8, with $\{\text{brightness, contrast, saturation, hue}\}$ 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.

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 pre-training epochs N_e is 200 (see Sec. 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 GPU.

IV. EXPERIMENTAL RESULTS

A. Pre-training Setup

Pre-training Dataset. We leverage image-mask pairs from the existing Inria building segmentation dataset [34], which contains 180 labeled RS images, each size of 5000×5000 and spatial resolution of 0.3 m. It provides pixel-level annotations, including the building and non-building categories. We cut the original samples into small patches of size 256×256 with no overlap and remove those without building regions. The resulting dataset contains more than 45k patch samples. We randomly split it into training (80%) and validation (20%) sets. We additionally obtain the co-registered 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 pre-training 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 pre-training framework.

B. Downstream Change Detection Setup

1) *Change detection datasets:* To evaluate the proposed pre-trained model, we conduct experiments on the following three change detection datasets:

LEVIR-CD [2]: a widely-used building change detection dataset. LEVIR-CD contains 637 pairs of bitemporal VHR RS images, each with a size of 1024×1024 and a spatial resolution of 0.5m. We follow the default train-valid-test split [2]. Each image is cropped with no overlap into small patches of size 256×256 . Finally, we obtain 7120/1024/2048 samples for training/validation/testing, respectively.

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

Guangzhou-CD [36]: a building change detection dataset that covers several suburb areas of Guangzhou City, China. It contains 19 VHR RGB image pairs (0.55m) with sizes ranging from 1006×1168 to 4936×5224 pixels. Similarly, the image pairs are cropped into 256×256 non-overlapping 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) *Change detection networks:* We employ a simple yet effective CD model [38] which consists of a Siamese feature extractor (deep 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 change detection networks employ a ResNet-18 backbone. We use the cross-entropy loss to optimize our CD model. Please refer to [38] for more details.

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

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.

Evaluation Metrics. The F1-score regarding the change category is used as the evaluation indices. Additionally, preci-

sion, recall, and Intersection over Union (IoU) of the change category are reported. These metrics are defined as follows:

$$\begin{aligned} F1 &= \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{IoU} &= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \end{aligned} \quad (14)$$

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 pre-training method, we make a comparison with the random initialization, ImageNet pre-training, in-domain supervised pre-training, and several self-supervised pre-training methods:

- *Random initialization.* All the parameters of the CD networks are initialized by the zero-mean normal distribution with a deviation of 0.02.
- *ImageNet pre-training.* The backbone (ResNet-18) of the CD model is initialized with the ImageNet pre-training.
- *In-domain supervised pre-training (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 pre-training 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.
- *SimSiam* [21]: A popular SSL method without using negative samples. SimSiam employs the stop-gradient and predictor to prevent model collapse.
- *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.
- *DenseCL* [58]: A dense contrastive learning 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.
- *CMC* [29]: A recent SSL method in RS that 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 multi-modal representation learning with image-mask pairs of our pre-training dataset.
- *SeCo* [16]: A very recent SSL method that is based on MoCo-v2 and further exploits the use of multi-temporal RS images as natural augmentations to learn seasonal invariant/variant representations. We implement it using the bitemporal images in our pre-training dataset.

For a fair comparison, we use our pre-training dataset to train these supervised/self-supervised methods, i.e., in-domain

supervision, SimSiam, MoCo-v2, DenseCL, 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 pre-trained weights as our SaDL does.

Each pre-trained model is evaluated by the performance of the downstream CD task. Tab. I reports the overall comparisons of our proposed method and other pre-trained models on the LEVIR-CD, 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 Tab. I, we can observe that the proposed method consistently outperforms the compared pre-trained models in terms of the F1-score across different data regimes on the three datasets. Specifically, our SaDL pre-training is significantly superior to random initiation. It is notable that we can achieve comparable or even better results with only 20% training data than the random initiation baseline using 100% training data. It indicates that our pre-training can effectively alleviate the labeled data insufficiency problem.

ImageNet pre-training 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 pre-training across different data conditions on the downstream CD task. For instance, Tab. I (a) shows that compared to ImageNet pre-training, 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%).

Comparison to in-domain supervision. Typically, supervised RS pre-training obtains in-domain discriminative representations by learning the mapping from RS images to labels. Tab. 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 pre-training in some data regimes. Note that we use the same amount of pre-training data for SaDL and in-domain supervision. The quantitative results indicate that our SaDL can learn more effective transferable features than the traditional approach. 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 pre-training may overfit patterns of a certain subdomain and lack transferability to another.

Comparison to other SSL pre-training. Tab. I also shows that our method consistently outperforms several SSL methods, including two popular image-level representation learning methods, one dense contrastive learning method, and two recent SSL methods in the RS domain. We can observe that the normal dense contrastive learning (DenseCL) can even be inferior to the image-level counterpart (MoCo) in many data conditions. It may 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 self-supervised learning 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 loses 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 pre-trained 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 pre-trained model achieves more accurate predictions than those of its counterparts.

Better convergence. We further give validation curves on the downstream CD task to show the transferability of each pre-trained 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 pre-training 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.

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 semantic discrimination (SD), and 3) learning consistent foreground features via background swap (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 pre-training 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.

Ablation on masked sampling (MS). Tab. 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 benefiting representation learning.

Ablation on semantic discrimination (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. Tab. 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 semantic discrimination 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 pre-training 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 benefiting the recognition of change of interest in the downstream CD task.

Ablation on background swap (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

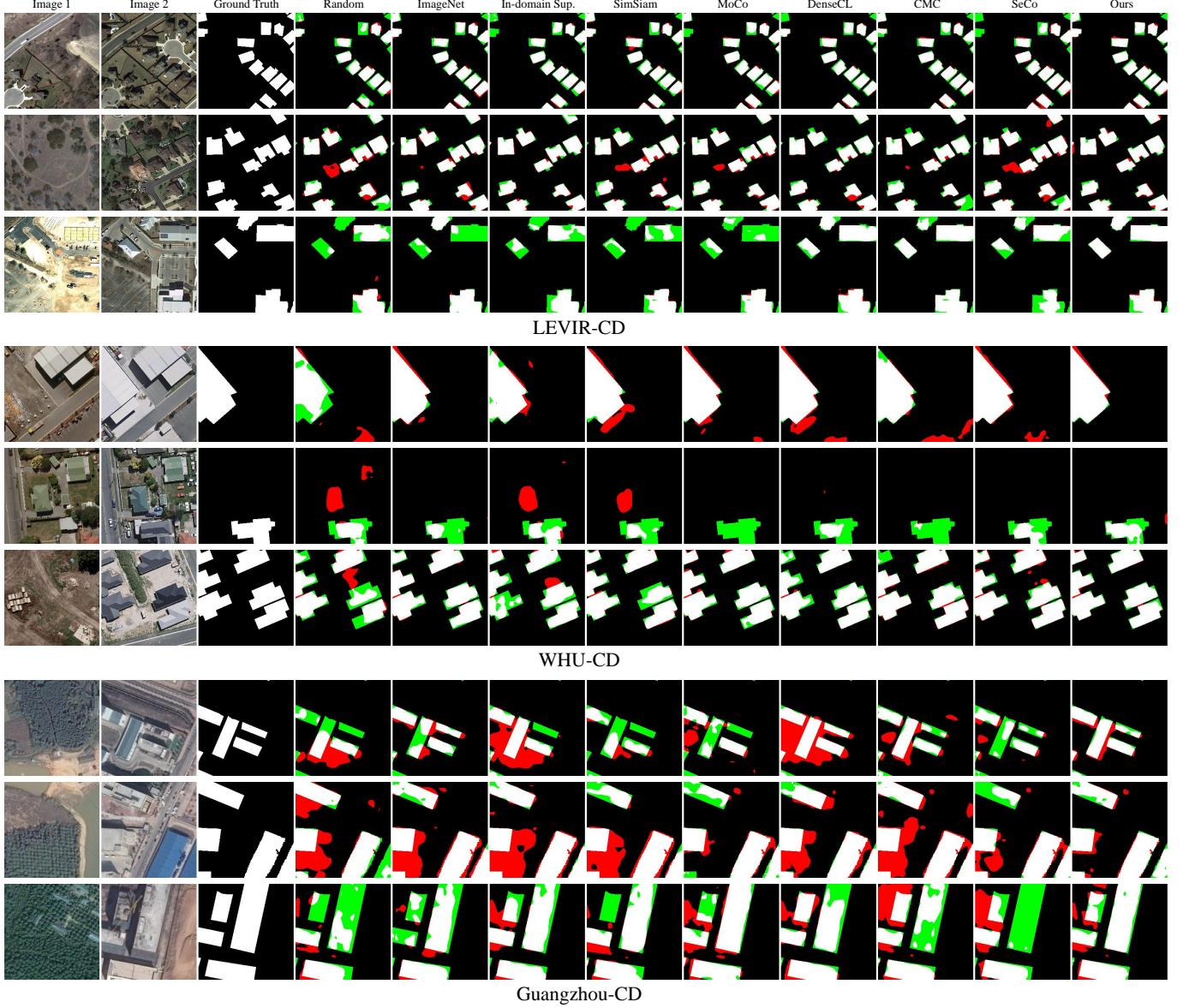


Fig. 4. Visualization comparison of the CD model fine-tuned from different pre-trained models on the LEVIR-CD/WHU-CD/Guangzhou-CD test sets. The CD model is trained under a 20% data regime. For a better view, white for true positives, black for true negatives, red for false positives, and green for false negatives.

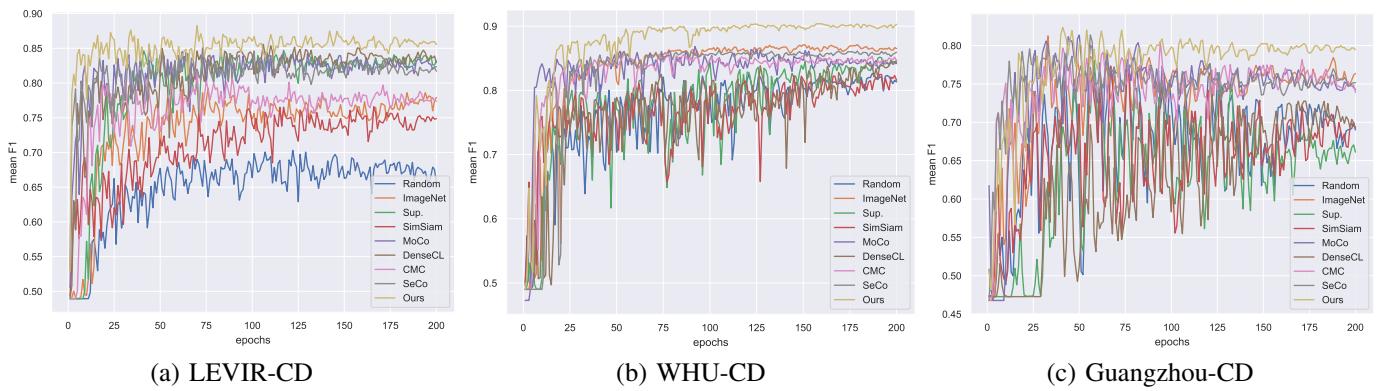


Fig. 5. Accuracy of CD models for each training epoch. The mean F1-score is reported.

TABLE I

OVERALL COMPARISONS OF DIFFERENT PRE-TRAINING 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.

(a) LEVIR-CD

	1%	5%	20%	100%
	Precision / Recall / F1 / IoU			
Random	49.83 / 13.56 / 21.32 / 11.93	73.54 / 34.58 / 47.05 / 30.76	83.00 / 77.76 / 80.30 / 67.08	90.50 / 84.94 / 87.63 / 77.98
ImageNet	30.80 / 20.11 / 24.33 / 13.85	81.40 / 54.35 / 65.18 / 48.35	89.59 / 82.26 / 85.76 / 75.08	91.29 / 85.25 / 88.16 / 78.83
In-domain Sup.	79.43 / 17.96 / 29.29 / 17.16	86.25 / 61.73 / 71.96 / 56.20	90.91 / 81.02 / 85.68 / 74.95	91.22 / 85.11 / 88.06 / 78.67
SimSiam [21]	71.90 / 13.29 / 22.44 / 12.64	82.46 / 44.41 / 57.73 / 40.58	88.97 / 76.46 / 82.24 / 69.84	91.47 / 84.78 / 88.00 / 78.57
MoCo-v2 [56]	58.15 / 20.30 / 30.09 / 17.71	74.78 / 71.20 / 72.95 / 57.42	88.78 / 80.45 / 84.41 / 73.03	91.59 / 85.06 / 88.21 / 78.90
DenseCL [58]	72.45 / 28.76 / 41.17 / 25.92	91.51 / 58.62 / 71.46 / 55.60	89.33 / 82.09 / 85.56 / 74.76	94.63 / 79.98 / 86.69 / 76.51
CMC [29]	79.43 / 10.42 / 18.42 / 10.15	84.45 / 56.40 / 67.63 / 51.09	90.30 / 77.94 / 83.67 / 71.92	90.73 / 86.59 / 88.61 / 79.55
SeCo [16]	68.63 / 29.01 / 40.78 / 25.62	84.74 / 61.43 / 71.23 / 55.31	84.34 / 83.70 / 84.02 / 72.44	91.03 / 85.68 / 88.27 / 79.01
Ours	82.38 / 42.52 / 56.09 / 38.98	88.35 / 72.16 / 79.44 / 65.90	89.58 / 85.05 / 87.25 / 77.39	90.91 / 86.66 / 88.74 / 79.75

(b) WHU-CD

	1%	5%	20%	100%
	Precision / Recall / F1 / IoU			
Random	37.03 / 41.76 / 39.25 / 24.42	67.99 / 64.40 / 66.14 / 49.41	70.07 / 72.68 / 71.35 / 55.46	85.82 / 83.20 / 84.49 / 73.15
ImageNet	31.37 / 33.55 / 32.42 / 19.35	80.12 / 67.92 / 73.52 / 58.12	87.82 / 76.87 / 81.98 / 69.46	91.11 / 86.14 / 88.55 / 79.46
In-domain Sup.	43.41 / 60.49 / 50.55 / 33.82	74.50 / 66.06 / 70.03 / 53.88	70.79 / 72.34 / 71.56 / 55.71	84.10 / 82.42 / 83.25 / 71.31
SimSiam [21]	37.86 / 47.38 / 42.09 / 26.65	60.37 / 73.03 / 66.10 / 49.36	70.89 / 76.11 / 73.41 / 57.99	86.30 / 83.14 / 84.69 / 73.45
MoCo-v2 [56]	71.36 / 57.44 / 63.65 / 46.68	76.45 / 67.21 / 71.53 / 55.68	85.30 / 76.36 / 80.58 / 67.48	91.36 / 84.69 / 87.90 / 78.41
DenseCL [58]	41.02 / 46.45 / 43.56 / 27.85	68.78 / 69.05 / 68.92 / 52.58	84.14 / 74.37 / 78.95 / 65.22	89.38 / 84.64 / 86.94 / 76.90
CMC [29]	50.89 / 45.32 / 47.94 / 31.53	73.69 / 67.54 / 70.49 / 54.42	86.28 / 74.54 / 79.98 / 66.64	89.96 / 85.47 / 87.66 / 78.03
SeCo [16]	67.39 / 41.92 / 51.68 / 34.85	78.81 / 66.52 / 72.15 / 56.43	83.93 / 76.43 / 80.00 / 66.67	89.29 / 86.93 / 88.09 / 78.72
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	93.10 / 87.05 / 89.98 / 81.78

(c) Guangzhou-CD

	1%	5%	20%	100%
	Precision / Recall / F1 / IoU			
Random	40.42 / 51.63 / 45.34 / 29.32	62.33 / 50.03 / 55.51 / 38.42	72.89 / 54.51 / 62.37 / 45.32	91.80 / 65.02 / 76.12 / 61.45
ImageNet	52.68 / 38.36 / 44.39 / 28.53	62.54 / 57.82 / 60.09 / 42.94	88.93 / 59.59 / 71.36 / 55.48	94.99 / 76.81 / 84.94 / 73.82
In-domain Sup.	36.54 / 51.99 / 42.92 / 27.32	66.53 / 42.04 / 51.52 / 34.70	69.82 / 60.43 / 64.79 / 47.91	88.96 / 73.11 / 80.26 / 67.03
SimSiam [21]	42.40 / 51.09 / 46.34 / 30.16	73.02 / 43.49 / 54.51 / 37.47	67.29 / 61.14 / 64.07 / 47.13	91.80 / 72.60 / 81.08 / 68.18
MoCo-v2 [56]	50.73 / 48.56 / 49.62 / 33.00	67.33 / 56.56 / 61.48 / 44.38	81.24 / 61.80 / 70.20 / 54.08	90.23 / 76.69 / 82.91 / 70.81
DenseCL [58]	37.06 / 51.16 / 42.98 / 27.37	58.69 / 40.78 / 48.12 / 31.69	77.36 / 57.88 / 66.21 / 49.49	93.19 / 71.36 / 80.83 / 67.82
CMC [29]	46.77 / 38.94 / 42.50 / 26.98	75.17 / 53.37 / 62.42 / 45.37	73.97 / 66.46 / 70.01 / 53.86	92.20 / 76.72 / 83.75 / 72.04
SeCo [16]	46.75 / 50.62 / 48.61 / 32.11	63.50 / 56.70 / 59.91 / 42.76	80.08 / 60.62 / 69.00 / 52.68	92.51 / 76.98 / 84.03 / 72.46
Ours	57.96 / 47.79 / 52.39 / 35.49	80.04 / 56.00 / 65.89 / 49.13	85.32 / 67.06 / 75.09 / 60.12	92.13 / 81.06 / 86.24 / 75.81

TABLE II

COMPUTATIONAL COMPLEXITY COMPARISON. WE REPORT THE NUMBER OF MODEL PARAMETERS (PARAMS.), FLOATING-POINT OPERATIONS PER SECOND (FLOPS) AND GPU INFERENCE TIME. THE INPUT IMAGE TO THE MODEL HAS A SIZE OF $256 \times 256 \times 3$.

Model	Params. (M)	FLOPs (G)	Time (ms)
SimSiam [21]	12.49	4.76	1.04
MoCo-v2 [56]	11.24	4.76	1.92
DenseCL [58]	11.69	4.76	2.66
CMC [29]	22.48	4.66	1.55
SeCo [16]	12.16	9.52	3.62
Ours	15.09	7.83	4.17

view 3. Quantitative results in Tab. III demonstrate that our BS further brings consistent improvements in the performance of the pre-trained CD model on the three CD datasets. It indicates that the learned foreground representations invariant to background changes benefit the downstream change detec-

tion for the land covers of interest. Our background swap 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. Tab. 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

Sampled point number. The number of sampled points N for each semantic category is an important hyperparameter. We

TABLE III

ABLATION STUDIES OF OUR SADL, INCLUDING 1) DENSE REPRESENTATION LEARNING VIA MASKED SAMPLING (MS), 2) LEARNING DISCRIMINATIVE FEATURES VIA SEMANTIC DISCRIMINATION (SD), AND 3) LEARNING CONSISTENT FOREGROUND FEATURES VIA BACKGROUND SWAP (BS). THE F1-SCORE ON THE LEVIR-CD/WHU-CD/GZ-CD TEST SET IS REPORTED. THE CD MODEL IS TRAINED UNDER THE DATA REGIME OF 5%.

	MS	SD	BS	LEVIR	WHU	GZ
Baseline	✗	✗	✗	60.69	64.53	48.24
+masked pool	✗	✗	✗	63.09	66.03	48.64
Ours	✓	✗	✗	70.21	71.90	54.15
Ours	✓	✓	✗	77.74	77.43	64.21
Ours	✓	✓	✓	79.44	80.34	65.89

TABLE IV

EFFECT OF THE THIRD VIEW. 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%.

	LEVIR-CD		WHU-CD		Guangzhou-CD	
	F1	IoU	F1	IoU	F1	IoU
base	77.74	63.58	77.43	63.17	64.21	47.29
+ natural	78.32	64.37	78.86	65.10	62.99	45.98
ours	79.44	65.90	80.34	67.15	65.89	49.13

test different $N \in \{1, 4, 16, 64\}$ to analyze its effect on the performance of the downstream fine-tuning performance. Tab. V reports the F1-score/IoU of the pre-trained 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 that 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.

Pre-training epochs number. The epochs number N_e of pre-training affects its fine-tuning performance on the downstream task. We train our SaDL using varying epochs numbers

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		WHU-CD		Guangzhou-CD	
	F1	IoU	F1	IoU	F1	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

TABLE VI

EFFECT OF THE NUMBER OF PRE-TRAINING EPOCHS. THE F1-SCORE AND IOU ON THE LEVIR-CD/WHU-CD/GUANGZHOU-CD TEST SETS ARE REPORTED. THE CD MODEL IS TRAINED UNDER THE DATA REGIME OF 5%.

epochs	LEVIR-CD		WHU-CD		Guangzhou-CD	
	F1	IoU	F1	IoU	F1	IoU
0	65.18	48.35	73.52	58.12	57.86	40.71
5	73.68	58.33	75.39	60.51	60.54	43.41
20	76.48	61.92	79.17	65.52	65.61	48.82
100	77.14	62.79	79.76	66.34	64.57	47.68
200	79.44	65.90	80.34	67.15	65.89	49.13
400	79.59	66.10	80.06	66.75	64.37	47.47

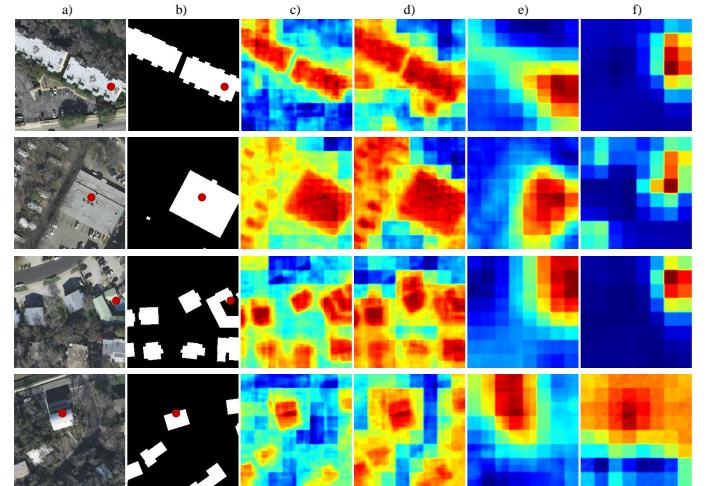


Fig. 6. Visualization of self-similarity map of our SaDL on the pre-training dataset. Given one selected point (viewed in red) in the a) image, we give visualization results of the ablation versions of our model, including f) baseline, e) ours with masked sampling (MS), d) ours with MS and semantic discrimination (SD), c) ours with MS, SD, and background swap (BS). The semantic mask b) is shown for a better view.

$N_e \in \{5, 20, 100, 200, 400\}$. Tab. VI reports the F1-score/IoU of the pre-trained CD model on the LEVIR-CD/WHU-CD/Guangzhou-CD datasets under a 5% data regime. Note that $N_e = 0$ denotes the ImageNet pre-training. We can observe a significant improvement in the F1-score even with a very small number of pre-training epochs (e.g., 5), compared to the baseline ($N_e = 0$). The performance of the pre-trained 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 pre-training with too many training epochs may overfit the pre-training 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

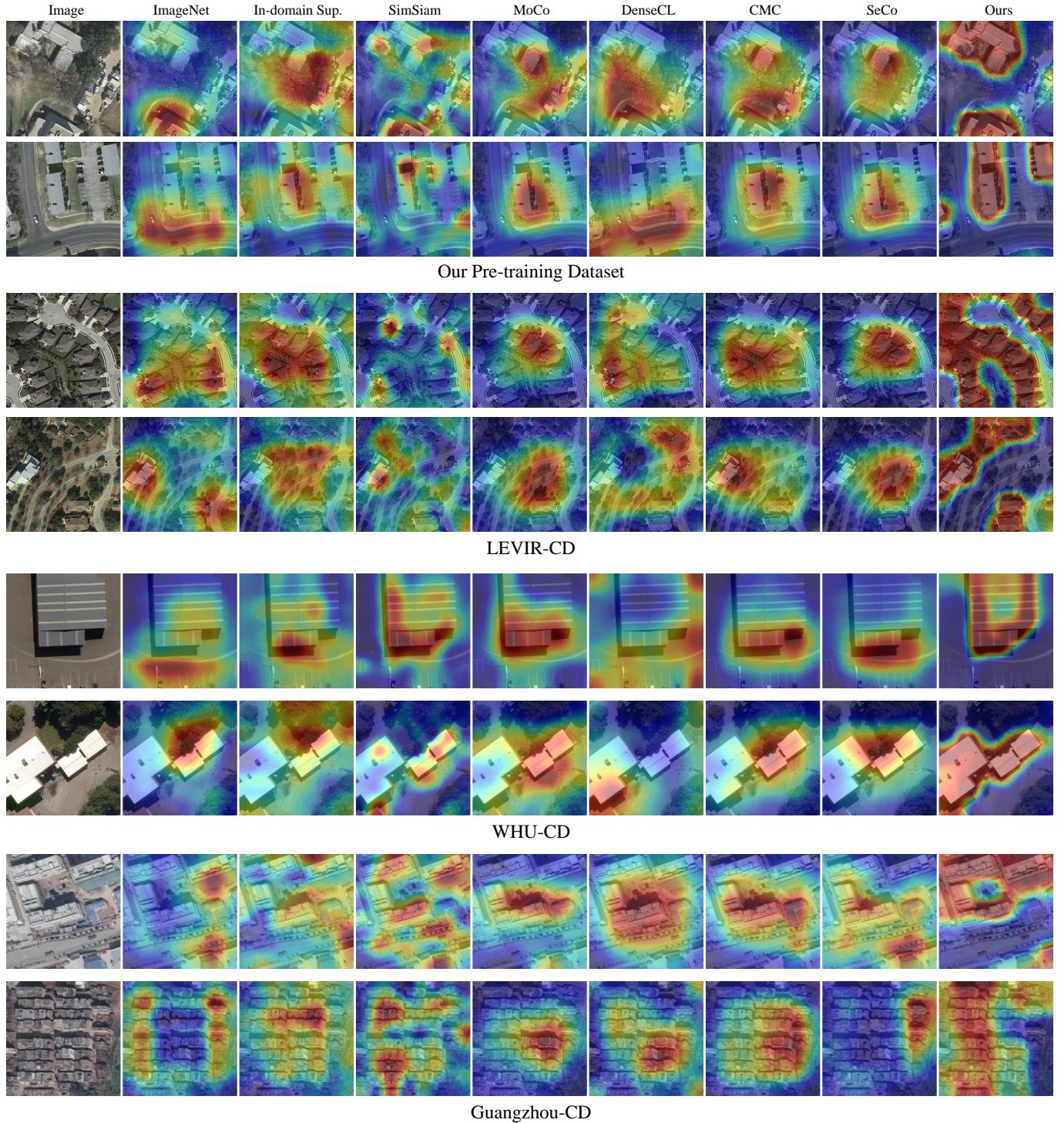


Fig. 7. Class activation maps of different pre-trained models. We test images from our pre-training dataset and the downstream LEVIR-CD/WHU-CD/Guangzhou-CD datasets.

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 $\mathbf{X} \in \mathbb{R}^{HW \times C}$ and one selected point feature $\mathbf{x} = \mathbf{X}[p] \in \mathbb{R}^C, p = (u, v)$, 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 pre-training 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 pre-trained models. To interpret what concepts are learned by the networks, we show the backbone visualization of each pre-trained 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 pre-trained models. Red denotes higher attention values and blue denotes lower values. We test images from our pre-training 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 pre-training images but also perform well on the downstream images with considerable domain gaps (e.g., differences in spatial resolution and object style) from the pre-training data. For instance, the high-density buildings (see the last row in Fig. 7) in the Guangzhou-CD dataset are hardly present in the pre-training dataset. It indicates that our proposed pre-training could extract transferable semantically meaningful representations with well

generalization ability.

V. DISCUSSION

We present semantic-aware representation learning for change detection in high-resolution remote sensing images. One major challenge of change detection 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 remote sensing community to learn a pre-trained model to benefit the downstream CD task. Considering the key of CD is to recognize the real changes and exclude other irrelevant changes, we incorporate such CD task priors in the pre-training 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 pre-training in a contrastive manner, which is different from traditional supervised pre-training that learns the mapping from the image to the label or other self-supervised learning methods that lack of semantic supervision. Compared to several existing pre-trained 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 pre-trained method can learn transferable discriminative features that benefit change recognition and incur better model generalization in the downstream CD task.

VI. CONCLUSION

In this paper, we proposed a semantic-aware pre-training method for remote sensing image change detection. We incorporate the semantic information into a self-supervised learning 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 background swap 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 pre-training, in-domain supervised pre-training, and several SSL pre-training 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 pre-training data.

Modern machine learning-based models typically suffer from spectral variabilities [84] in remote sensing 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 unconcerned 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 pre-training 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 pre-training backbones, e.g., more kinds of CNN structures and even other vision transformer architectures. 3) extending our semantic-aware pre-training to distinguish more land cover types, which may benefit the multi-class semantic change detection task.

REFERENCES

- [1] A. SINGH, “Review article digital change detection techniques using remotely-sensed data,” *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989. [Online]. Available: <https://doi.org/10.1080/01431168908903939>
- [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. 10, p. 1662, 2020.
- [3] L. Bruzzone and D. Prieto, “Automatic analysis of the difference image for unsupervised change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.
- [4] P. P. de Bem, O. A. de Carvalho Junior, R. F. Guimarães, and R. A. T. Gomes, “Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks,” *Remote Sensing*, vol. 12, no. 6, p. 901, 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.
- [6] L. Bruzzone and F. Bovolo, “A novel framework for the design of change-detection systems for very-high-resolution remote sensing images,” *Proceedings of the IEEE*, vol. 101, no. 3, pp. 609–630, 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 Sensing*, vol. 12, p. 1688, 5 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, 2019.
- [9] F. I. Diakogiannis, F. Waldner, and P. Caccetta, “Looking for change? roll the dice and demand attention.”
- [10] M. Zhang and W. Shi, “A feature difference convolutional neural network-based change detection method,” *TGRS*, pp. 1–15, 2020.
- [11] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, “Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2021.
- [12] S. Saha, F. Bovolo, and L. Bruzzone, “Building change detection in vhr sar images via unsupervised deep transcoding,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2020.
- [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, 2020.
- [14] J. Liu, K. Chen, G. Xu, X. Sun, M. Yan, W. Diao, and H. Han, “Convolutional neural network-based transfer learning for optical aerial images change detection,” vol. 17, pp. 127–131, 2020.
- [15] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, “An Empirical Study of Remote Sensing Pretraining,” *arXiv:2204.02825 [cs]*, Apr. 2022.
- [16] O. Manas, A. Lacoste, X. G. i Nieto, D. Vazquez, and P. Rodriguez, “Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 9394–9403. [Online]. Available: <https://doi.ieee.org/10.1109/ICCV48922.2021.00928>
- [17] M. Leenstra, D. Marcos, F. Bovolo, and D. Tuia, “Self-supervised pre-training enhances change detection in sentinel-2 imagery,” *Part of the Lecture Notes in Computer Science book series (LNCS, volume 12667)*, 2021, Jan. 2021.
- [18] P. Akiva, M. Purri, and M. Leotta, “Self-Supervised Material and Texture Representation Learning for Remote Sensing Tasks,” *CoRR*, Dec. 2021, arXiv: 2112.01715. [Online]. Available: <http://arxiv.org/abs/2112.01715>
- [19] E. Sanchez, M. Serrurier, and M. Ortner, “Learning disentangled representations of satellite image time series,” *CoRR*, Mar. 2019.
- [20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.
- [21] X. Chen and K. He, “Exploring simple siamese representation learning,” *CoRR*, vol. abs/2011.10566, 2020. [Online]. Available: <https://arxiv.org/abs/2011.10566>
- [22] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, “Geography-aware self-supervised learning,” *ICCV*, 2021.
- [23] C. Tao, J. Qia, G. Zhang, Q. Zhu, W. Lu, and H. Li, “TOV: The Original Vision Model for Optical Remote Sensing Image Understanding via Self-supervised Learning,” *arXiv:2204.04716 [cs]*, Apr. 2022.
- [24] 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 Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.
- [25] V. Stojnic and V. Risojevic, “Self-supervised learning of remote sensing scene representations using contrastive multiview coding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021.
- [26] A. M. Swope, X. H. Rudelis, and K. T. Story, “Representation learning for remote sensing: An unsupervised sensor fusion approach,” Aug. 2021.
- [27] H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, “Contrastive Self-Supervised Learning With Smoothed Representation for Remote Sensing,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.
- [28] C. Agastya, S. Ghebremusse, I. Anderson, C. Reed, H. Vahabi, and A. Todeschini, “Self-supervised contrastive learning for irrigation detection in satellite imagery,” *CoRR*, vol. abs/2108.05484, 2021. [Online]. Available: <https://arxiv.org/abs/2108.05484>
- [29] K. Cha, J. Seo, and Y. Choi, “Contrastive multiview coding with electro-optics for SAR semantic segmentation,” *CoRR*, vol. abs/2109.00120, 2021. [Online]. Available: <https://arxiv.org/abs/2109.00120>

- [30] W. Li, H. Chen, and Z. Shi, "Semantic segmentation of remote sensing images with self-supervised multitask representation learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 6438–6450, 2021.
- [31] H. Li, Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao, "Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [32] M. Tarasiou and S. Zafeiriou, "Embedding Earth: Self-supervised contrastive pre-training for dense land cover classification," *arXiv:2203.06041 [cs]*, Mar. 2022.
- [33] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.
- [34] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2017, Fort Worth, TX, USA, July 23-28, 2017*. IEEE, 2017, pp. 3226–3229.
- [35] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019.
- [36] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "Semicdnet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2020.
- [37] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS*, vol. 166, pp. 183–200, 2020.
- [38] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021.
- [39] X. Zhang, W. Yu, and M.-O. Pun, "Multilevel deformable attention-aggregated networks for change detection in bitemporal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.
- [40] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Jun. 2021.
- [41] T. Liu, M. Gong, D. Lu, Q. Zhang, H. Zheng, F. Jiang, and M. Zhang, "Building Change Detection for VHR Remote Sensing Images via Local–Global Pyramid Network and Cross-Task Transfer Learning Strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, Nov. 2021.
- [42] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.
- [43] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, "Difference Enhancement and Spatial–Spectral Nonlocal Network for Change Detection in VHR Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, Dec. 2021.
- [44] S. Hafner, Y. Ban, and A. Naselli, "Urban Change Detection Using a Dual-Task Siamese Network and Semi-Supervised Learning," *arXiv*, Apr. 2022.
- [45] W. G. C. Bandara and V. M. Patel, "Revisiting Consistency Regularization for Semi-supervised Change Detection in Remote Sensing Images," *arXiv:2204.08454 [cs, eess]*, Apr. 2022.
- [46] V. Ruzicka, S. D'Aronco, J. D. Wegner, and K. Schindler, "Deep active learning in remote sensing for data efficient change detection," in *Proceedings of MACLEAN: MACHINE Learning for EAth ObservatioN Workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2020), Virtual Conference, September 14-18, 2020*, ser. CEUR Workshop Proceedings, T. Corpetti, D. Ienco, R. Interdonato, M. Pham, and S. Lefèvre, Eds., vol. 2766. CEUR-WS.org, 2020. [Online]. Available: <http://ceur-ws.org/Vol-2766/paper2.pdf>
- [47] X. Wang, P. Du, D. Chen, S. Liu, W. Zhang, and E. Li, "Change detection based on low-level to high-level features integration with limited samples," vol. 13, pp. 6260–6276, 2020.
- [48] J. Li, X. Huang, and X. Chang, "A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 1–17, may 2020.
- [49] H. Sahbi, S. Deschamps, and A. Stoian, "Active learning for interactive satellite image change detection," *arXiv*, Oct. 2021.
- [50] V. Risojevic and V. Stojnic, "The Role of Pre-Training in High-Resolution Remote Sensing Scene Classification," *arXiv:2111.03690 [cs]*, Nov. 2021, arXiv: 2111.03690. [Online]. Available: <http://arxiv.org/abs/2111.03690>
- [51] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, "Training general representations for remote sensing using in-domain knowledge."
- [52] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical Knowledge-Driven Representation Learning for Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Oct. 2021.
- [53] H. Chen, Y. Zao, L. Liu, S. Chen, and Z. Shi, "Semantic decoupled representation learning for remote sensing image change detection," *arXiv*, Jan. 2022.
- [54] X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive."
- [55] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607. [Online]. Available: <http://proceedings.mlr.press/v119/chen20j.html>
- [56] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning."
- [57] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html>
- [58] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3023–3032.
- [59] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 3733–3742. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00393>
- [60] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bo-

- janowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.
- [61] N. Jean, S. Wang, A. Samar, G. Azzari, D. B. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 3967–3974.
- [62] 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 Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.
- [63] B. Peng, Q. Huang, and J. Rao, "Spatiotemporal contrastive representation learning for building damage classification," in *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2021, Brussels, Belgium, July 11-16, 2021*. IEEE, 2021, pp. 8562–8565.
- [64] P. Jain, B. Schoen-Phelan, and R. Ross, "Self-Supervised Learning for Invariant Representations from Multi-Spectral and SAR Images," *arXiv*, May 2022.
- [65] Y. Chen and L. Bruzzone, "Self-supervised change detection in multiview remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
- [66] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.
- [67] Y. Chen and L. Bruzzone, "Self-supervised Remote Sensing Images Change Detection at Pixel-level," *arXiv*, May 2021.
- [68] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2021.
- [69] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
- [70] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [71] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2021.
- [72] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.
- [73] 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.
- [74] J. Wu, B. Li, Y. Qin, W. Ni, H. Zhang, R. Fu, and Y. Sun, "A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 105, p. 102615, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0303243421003226>
- [75] J. Qu, Y. Xu, W. Dong, Y. Li, and Q. Du, "Dual-branch difference amplification graph convolutional network for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
- [76] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [77] W. G. C. Bandara and V. M. Patel, "A Transformer-Based Siamese Network for Change Detection," *arXiv*, Jan. 2022.
- [78] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization."
- [79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.
- [80] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017, pp. 936–944.
- [81] 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, 2018.
- [82] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 776–794. [Online]. Available: https://doi.org/10.1007/978-3-030-58621-8_45
- [83] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 111–119. [Online]. Available: <https://doi.ieee.org/10.1109/CVPRW50498.2020.00020>
- [84] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2019.