Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection

Tian Wei, Youfa Liu†, Rui Zhao†

School of Computer Science, Wuhan University, Wuhan, 430072, China
Faculty of Electrical Engineering and Computer Science, Ningbo University, 315211, China

Abstract

Significant progress has been made in remote sensing image change detection due to the rapid development of deep learning techniques. Convolutional neural networks (CNNs) have become foundational models in this field. Previous works on remote sensing image change detection has utilized domain adaptation methods, achieving promising predictive performance. However, the transferable knowledge between source and target domain has not been fully exploited. In this paper, we propose a novel cross-domain contrastive learning approach for remote sensing image change detection, which correlates source and target domain using contrastive principles. Specifically, we introduce a transferable cross-domain dictionary learning scheme where a shared dictionary between the source and target domains generates sparse representations. Based on these representations, we compute attention weights and propose an attention-weighted contrastive loss to enhance knowledge transfer between source and target domains. Experiments demonstrate the effectiveness of the proposed methods on public remote sensing image change detection datasets.

Keywords: Change detection; Domain adaptation; Contrastive learning; Dictionary learning; Attention mechanism

1. Introduction

Change detection is one of the major topics in remote sensing (RS) [1], aiming to identify changes in the same area over different periods. The types of changes detected can vary based on the purpose, such as building alterations, road modifications, farmland transformations, and environmental changes. In recent years, with the increasing volume of remote sensing data, there is a growing demand for more accurate and generalized change detection models. However, a significant challenge arises due to the difficulty in acquiring annotated data, as manual pixel-level annotation is prohibitively expensive. Consequently, most remote sensing images remain unlabeled. Furthermore, there are considerable domain distribution differences between various remote sensing datasets, caused by external factors such as sensor types, illumination conditions, and temporal variations. In summary, the lack of sufficient annotated datasets and the limited

†Corresponding author: Youfa Liu (Email: liuyfa1991@whu.edu.cn; ORCID: 0000-0002-3540-5775), Rui Zhao (Email: zhaorui@nbu.edu.cn; ORCID: 0000-0002-9577-7714)
generalization ability of trained models present a challenge. Effectively utilizing these unlabeled data becomes critical, posing higher requirements for the change detection task.

Change detection methods have provided many insights into handling unlabeled data. Most approaches employ Semi-Supervised Learning (SSL) to solve the problem, which necessitates a small number of annotated images in the dataset. However, scenarios where the dataset lacks annotations have received little attention from researchers. Based on the above reason, we propose an unsupervised domain adaptation framework to utilize the unlabeled remote sensing images, called Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection (DCCD). This framework can suppress the negative impact of domain differences between datasets and employs a weighted contrastive learning module to enhance representation. This framework is divided into two stages: the mixed domain stage and the alignment stage. In the first stage, DCCD adopts Dual Soft Paste (DSP) strategy to construct the mixed domain by copying image patches from source domain to both source and target images, which builds a bridge between the two domains. In the second stage, alignment is achieved in terms of network output and features by adding constraints. This narrows the feature distance between the two domains, thereby improving the network’s performance in the target domain.

In the experiment, we utilized two datasets to evaluate the performance of our method. We found that our framework, when combined with other state-of-the-art methods, achieved better performance than the original method on a new unlabeled dataset. In addition, we also conducted ablation experiment for several hyper-parameters. These experiments demonstrate that our framework improves the generalization ability of method. Thus, our framework enhances baseline performance in the unsupervised cross-domain change detection task.

In summary, the main contributions of our work can be summarized as follows.

1) We design a Contrastive Learning Module for cross-domain unsupervised change detection task, which reduces the distance between the closest pairs of images from two domains based on their similarity.

2) To further eliminate differences between two domains, dictionary learning is introduced and it learns the attention between different image features.

3) Extensive experiments on two datasets demonstrate that our framework can obtain better performance than original method, which shows that our framework has better generalization ability.

The remainder of the paper is organized as follows. Section 2 summarizes related work for this study. Section 3 provides a detailed explanation of our work. Section 4 presents the extensive experiment results about this study and analyzes them. Finally, conclusion is drawn in Section 5.

2. Related Work

The application of unsupervised domain adaptation is commonly existent in real world. And it is gradually becoming important.

Change detection is one of the hottest topics in remote sensing image processing. Due to the increasing data of remote sensing images, the application of change detection becomes more important. It has a wide range of applications, such as environmental monitoring, military monitoring, intelligent traffic monitoring, disaster monitoring and so on. Generally, input of change detection model is a pair of images in the same area, those are captured at different times. The output is change image, which has the same size as input image and the value of each pixel
Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection

represents whether the corresponding position in the original image has changed. Samadi et al. [2] propose a supervised deep learning change detection method which combines bitemporal image with morphological images. The output of it is fed to DBN as good data and then getting good performance in experiment. SNUNet [3] proposes densely connected network for change detection, which utilizes shallow layer feature to obtain fine-grained localization information. BIT [1] uses transformer to model the spatiotemporal context of remote sensing images, and it improves performance at the cost of increasing computational complexity. ChangeFormer [4] unifies hierarchically structured transformer encoder with Multi-Layer Perceptron (MLP) decoder in a Siamese network architecture to obtain information in different scale, which helps it to get higher result. WNet [5] proposes a W-shaped dual-Siamese branch hierarchical network, which extracts local and global long-range features simultaneously. It has more flexible receptive fields to detect the regions, but the parameter size becomes too large. Zhengshun Du et al. [24] propose a framework, MTCDN, that fuses image translation and change detection tasks for optical and synthetic aperture radar (SAR) images. MTCDN eliminates main heterogeneous features through image translation with cyclic structure, and establishes a compact multitask structure, making it efficient and precise. SEIFNet [25] extracts multi-scale graphs via siamese networks, incorporates the spatiotemporal difference enhancement module (ST-DEM) to learn target-level global changing information and fine-grained local content, and finally aggregates multi-scale features using adaptive context fusion module (ACFM).

**Domain adaptation** is a branch of transfer learning that aims to address the problem of inconsistent distributions between two datasets. Domain shift is a common issue among various datasets. Reducing the domain shift by mapping source domain and target domain to the same feature space is the central idea of domain adaptation. [6] investigates challenge of unlabeled target data. It introduces generative adversarial net (GAN) to bypass the dependence on the source data and then uses predictor to acquire results. CLDA [7] introduces two contrastive losses, class-wise contrastive learning and instance-level contrastive alignment, to bridge the gap between labeled source domain and unlabeled target domain in classification.

**Contrastive learning** is a special unsupervised learning. The main idea of it is maximizing the similarity between relevant samples while minimizing the similarity between irrelevant ones. How to define the positive and negative pair, feature extraction and loss function are mainly ideas that need to be designed. [8] utilizes different data augmentations to easily obtain positive pair and negative pair. Otherwise, it adds the non-linear transform operation into framework to improve the representation performance. DCL [9] achieves decoupling of positive and negative pairs by removing the positive pair in denominator. This method eliminates negative-positive-coupling effect and especially improves the performance in small batch size. CLSA [10] integrates different data augmentations to achieve stronger augmentation and weak augmentation. And then using weak augmentation to supervise stronger augmentation in training by minimizing the distribution divergence between strong and weak images. ProCo [11] wants to alleviate the long tail problem in large dataset through modeling probabilistic distribution for every class.

**Dictionary learning** is a decomposition method, whose goal is to extract the essence of the object and use a dictionary to represent the characteristics of them. Intuitively, dictionary learning will decompose the original sample \( Y \in \mathbb{R}^{m \times n} \) into dictionary matrix \( D \in \mathbb{R}^{m \times K} \) and sparse matrix \( X \in \mathbb{R}^{K \times n} \), where \( K \) is the number of atoms in dictionary and \( Y = DX \). K-SVD [12] is a classic Dictionary learning method, which uses Singular Value Decomposition (SVD) to generate and update dictionary. Deep Dictionary Learning [13] combines deep learning with dictionary learning. The deep learning part extracts features and learns the filter. The dictionary learning part is used to learn features by decomposition. These two important patterns help
Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection

each other in classification and clustering task. DDLCN [14] is also a method combined these two paradigms, which is concentrated on image-recognition tasks with limited data. DDLCN proposes compound dictionary learning and coding layers to replace the convolutional layers in CNN. The new layers can transmit more fine-grained information to next layer through the activated atoms. That makes DDLCN achieves competitive results. [13] introduces “Bridge Integration”, a method to integrate single-cell datasets across modalities using a multionic dataset as a molecular bridge. To combine dictionary learning with sketching techniques to improve computational scalability. Dic-Attn [16] is developing attention algorithms by using sparsity (namely dictionary learning), which decompose the input feature, and update the dictionary and sparse representation respectively, finally merge them again to acquire the attention map.

3. The Proposed Approach

In this section, we introduce the detailed information of our framework. First, we present the structure of the proposed framework, as shown in Fig. 1. Next, the content of module is described in detail. Finally, we elaborate on the loss function.

![Figure 1](image)

**Figure 1**: The overall structure of DCCD. Arrows of different colors represent data flows from different domains. Due to the domain difference between source domain and target domain, this work first uses DSP (double black arrows) to construct the mixed domain to reduce it. After that, input the original and mixed images into network for alignment. Finally, teacher network is updated by using exponential moving average. Student network is used in test phase.

3.1. Overall Structure

The overall structure of DCCD is illustrated in Fig.1. It employs a Student-Teacher network to mitigate the negative impact of domain differences between source domain and target domain. DCCD is built upon the Average Teacher Framework [17] and integrates the Dual Soft Paste strategy, a contrastive learning module and a dictionary attention-weighted module. In the first stage, given annotated source domain bitemporal image pair \((I_{S-1}, I_{S-2}) \in S\) and target domain bitemporal image pair \((I_{T-1}, I_{T-2}) \in T\) with no annotation, where \(S\) and \(T\) represent source domain and target domain respectively. The DSP strategy randomly selects one bitemporal image pair \((I_{ST-1}, I_{ST-2}) \in S\) as template image pair from source domain. These three pairs of image construct the mixed domain. Specifically, the change area inside template image pair is cropped
and pasted into source domain image pair and target image pair. Thus, the source-mixed image pair \((I^s_{-M-1}, I^s_{-M-2})\) and target-mixed image pair \((I^t_{-M-1}, I^t_{-M-2})\) are obtained. For second stage, we input these image pairs into student network for learning. And then, using the parameters of student network to perform exponential moving average on the teacher network. With the constraint of loss function, the features of two domain are aligned and domain difference is reduced.

### 3.2. Copy-Paste Strategy

To reduce the significant domain difference between source domain and target domain, copy-paste strategy \([18]\) is introduced to construct the mixed domain. The copy-paste method copies the changed area from the template image pair to source and target domain image pairs. The reason for copying the changed areas is to maximize their proportion in the images. That alleviates the problem of relatively low overall proportion of changed regions in various datasets.

In the real implementation, first randomly select a bitemporal template pair \((I^{ST}_{-1}, I^{ST}_{-2}) \in S\) from source domain. Mask \(M\) is used to select the part which should be pasted. It is shown as follows:

\[
M(i, j) = \begin{cases} 
1, & \text{if } Y_{temp}(i, j) = 1 \\
0, & \text{otherwise} 
\end{cases}
\]  

(1)

where \(Y_{temp}\) represents the change detection ground truth of the template image pair. When the area is changed area, the value of \(M\) is 1, which means this area is used as copy-paste area. On the contrary, when the area is not a changed area, the value of \(M\) is 0, that indicates the area is not used as copy-paste area.

Then, the mixed domain image pair can be obtained with the help of \(M\). The source-mixed image pair \((I^s_{-M-1}, I^s_{-M-2})\) can be expressed as follows:

\[
\begin{align*}
I^s_{-M-1} &= \beta M \odot I^{ST}_{-1} + (1 - \beta M) \odot I^s_{-1} \\
I^s_{-M-2} &= \beta M \odot I^{ST}_{-2} + (1 - \beta M) \odot I^s_{-2}
\end{align*}
\]

(2)

where \(\odot\) indicates multiplication at corresponding position. And \(\beta \in (0, 1)\) represents transparency. It ensures that the mixed images retain the original image information. Similarly, the target-mixed image pair can be obtained by:

\[
\begin{align*}
I^t_{-M-1} &= \beta M \odot I^{ST}_{-1} + (1 - \beta M) \odot I^t_{-1} \\
I^t_{-M-2} &= \beta M \odot I^{ST}_{-2} + (1 - \beta M) \odot I^t_{-2}
\end{align*}
\]

(3)

As a result, two pairs of mixed domain image is obtained. In these images, the copy-paste area contains information from two domains, which constructs bridge between two domains. In next stage, the mixed images are used for output alignment.

### 3.3. Mean Teacher Framework

In the training process, this work uses two networks with the same structure, which can be easily replaced by various detection models. According to the average teacher framework, they are referred to as the teacher model and the student model. During the training, the data will be input into student model for learning, and then the exponential moving average (EMA) will be used to update the parameters in teacher model.
For this framework, image pairs from source domain, target domain and mixed domain are all input into student model for learning. As shown in Fig[1] the source domain image pair \((I_{S-1}, I_{S-2})\) is input into student model, and the output is a binary prediction image \(O_S\). Combined with ground truth, the change detection loss can be calculated as:

\[
\mathcal{L}_{CD} = \mathcal{L}(O_S, Y_S)
\]

where \(Y_S\) represents ground truth of source domain, and \(\mathcal{L}(\cdot)\) follows the loss function of change detection network.

Similarly, the source-mixed image pair \((I_{S-M-1}, I_{S-M-2})\) is input into student model \(f_\theta\), and the prediction output denotes as \(O_{S-M} = f_\theta(I_{S-M-1}, I_{S-M-2})\), the loss is shown as follows:

\[
\mathcal{L}_{CD-M}\|f\| = \mathcal{L}(O_{S-M}, Y_{S-M}) \odot M + \mathcal{L}(O_{S-M}, Y_S) \odot (1 - M)
\]

where \(Y_{S-M}\) represents ground truth of source-mixed domain, which can be obtained by combining annotation information from source domain image \(Y_S\) and annotation information from template image \(Y_{temp}\) with copy-paste method, namely Eq[2].

In addition, for the target domain image pair \((I_{T-1}, I_{T-2})\), it is fed into teacher model \(f'_\theta\), then we can obtain the pseudo-label as follows:

\[
\hat{Y}_T = f'_\theta(I_{T-1}, I_{T-2})
\]

where \(\hat{Y}_T\) denotes pseudo-label of target domain. Meanwhile, input target-mixed image pair \((I_{T-M-1}, I_{T-M-2})\) into student model, the output is corresponding prediction image \(O_{T-M}\). By using \(\hat{Y}_T\) and \(O_{T-M}\), the consistency loss can be obtained:

\[
\mathcal{L}_{cons} = \mathcal{L}(O_{T-M}, Y_{T-M}) \odot M + \mathcal{L}(O_{T-M}, \hat{Y}_T) \odot (1 - M)
\]

where \(Y_{T-M}\) represents ground truth of target-mixed domain, which is a binary image obtained by merging \(\hat{Y}_T\) and template image ground truth \(Y_{5T}\) via copy-paste method, like \(Y_{S-M}\). Besides the output alignment above, this study also contains feature alignment. Source-mixed image pair \((I_{S-M-1}, I_{S-M-2})\) and target-mixed image pair \((I_{T-M-1}, I_{T-M-2})\) have the same copy-paste patch, so we hope the features extracted from their corresponding area are as consistent as possible, which implicitly reduces the differences between two domains. For this issue, Maximum Mean Discrepancy (MMD) [19] is a suitable method introduced to align two domains, which is widely used in field of domain adaptation to measure the similarity between two distributions. MMD judges whether two distributions belong to the same distribution based on the statistical characteristics of samples drawn from each. A easily computable form of MMD is expressed as:

\[
MMD[\mathcal{F}, p, q] = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)])
\]

where \(\mathcal{F}\) means a class of function, \(p\) and \(q\) are two distributions. The function class should ideally be rich enough to uniquely determine whether \(p\) equals \(q\), while also being sufficiently constrained to provide useful finite sample estimates. Reproducing kernel Hilbert space (RKHS) [20] is the function that meets these two criteria. Therefore, when we select RKHS, Eq[8] can also be written as:

\[
MMD[\mathcal{F}, p, q] = \|\mu[p] - \mu[q]\|_H
\]

where \(\mu[p]\) and \(\mu[q]\) are the mean element of RKHS of \(\mu[p]\) and \(\mu[q]\).
where \( \mu(\cdot) \) denotes the kernel mean embedding and \( \mathcal{H} \) denotes RKHS. Using it to align the copy-paste patch features can be expressed as follows:

\[
L_{\text{paste}} = \| \mu(f_e(I_{S-M-1}, I_{S-M-2}) \odot M) - \mu(f_e(I_{T-M-1}, I_{T-M-2}) \odot M) \|_{\mathcal{H}}^2
\]  

(10)

where \( f_e \) represents the feature extractor of student model \( f_\theta \). Besides the copy-paste area should maintain consistency, the feature of the whole image from two mixed domain should also be consistent. This study still uses MMD to calculate the global loss:

\[
L_{\text{global}} = \| \mu(f_e(I_{S-M-1}, I_{S-M-2})) - \mu(f_e(I_{T-M-1}, I_{T-M-2})) \|_{\mathcal{H}}^2
\]  

(11)

In summary, we can consider the loss of alignment stage as:

\[
L_1 = L_{\text{CD}} + L_{\text{CD-soft}} + L_{\text{cons}} + \lambda_{\text{feature}}(L_{\text{paste}} + L_{\text{global}})
\]  

(12)

where \( \lambda_{\text{feature}} \) is a hyper-parameter to balance different losses, default to 0.005. Our intention is to balance the final values of each loss, as we added a small hyper-parameter due to the larger magnitude of the last set of losses. For the other several losses, their magnitudes are similar, so we did not add weights to them.

### 3.4. Contrastive Learning Module

In addition to above loss function, we also use contrastive learning to reduce the domain difference. The main idea of contrastive learning is to increase the similarity of positive pair while reducing the similarity of negative pairs.

In this work, the input of contrastive learning module is a batch of prediction images obtained from the student model. For example, if the batch size is \( N \) (default \( N = 3 \)), then \( N \) source image pairs and \( N \) target image pairs are used as input. As a result, \( 2N \) prediction images are generated by the student model. Both positive pair and negative pair are generated in these \( 2N \) images.

Firstly, calculating the similarity between source domain prediction image \( [x_1, x_2, x_3] \) and target domain prediction image \( [x_4, x_5, x_6] \). Secondly, make each source domain prediction image and another target domain prediction image that has the highest similarity to it become a positive pair, and the rest is negative pair. Finally, the contrastive loss is described as follows:

\[
L_{\text{cont}} = -\log \frac{\exp(sim(x_i, x_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(sim(x_i, x_k)/\tau)}
\]  

(13)

where \( sim(\cdot) \) means the similarity, \( \tau \) is temperature coefficient, \( \mathbb{1}_{[k \neq i]} \) means that it is 1 when \( k \neq i \), and it is 0 when \( k = i \). Besides, \( (x_i, x_j) \) represents a positive pair, \( (x_i, x_k) \) where \( k \neq i \) denotes other negative pairs. And only one source prediction image and one target prediction image can form a positive pair. The more similarity of positive pair is and the less similarity of negative pairs are, the smaller \( L_{\text{cont}} \). So the result is that source domain and target domain become closer and closer during the training.

### 3.5. Dictionary Attention-Weighted Module

To further enhance contrastive learning, we incorporate dictionary learning as an attention weight. As it mentioned in Section 2, the goal of dictionary learning is to extract the most essential features of objects as the dictionary and then use the sparse representation to reconstruct the image. Notably, the dictionary \( D \) is shared between the source and target domains, making
Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection

dictionary learning equivalent to mapping both domains to the same feature space, which further reduces the domain differences.

For the dictionary attention-weighted module, we consider the $2N$ prediction images as the original sample $X \in \mathbb{R}^{2N \times d}$, where $d = W \times H \times C$ is the length of the vector flattened by prediction image, namely the product of prediction image width, height and number of channels. Sometimes errors may occur due to $d$ being too long, and thus we also can reshape the size of $X$ to any more convenient form. Original sample $X$ is determined, the learnable dictionary $D \in \mathbb{R}^{2N \times K}$ and sparse representation $\alpha \in \mathbb{R}^{K \times d}$ ideally should be $X = D \cdot \alpha$, where we use complete dictionary so $K$ equals to the number of row of $X$. There is a dictionary loss:

$$L_{\text{dic}} = \|X - Da\|_2^2 + \|\alpha\|_1 \quad (14)$$

The first term ensures that $D$ and $\alpha$ can reconstruct $X$ as much as possible. And the second term ensures the sparsity of $\alpha$. Finally, we use $\alpha$ to calculate the attention between two domain prediction images. It is notable that after learning the dictionary, it is still necessary to reshape the size of $\alpha$ to $2N \times d$, through the similarity calculation, the attention matrix $W \in \mathbb{R}^{2N \times 2N}$ is acquired:

$$W = \frac{\alpha \alpha^T + 1}{2\alpha_{\text{norm}}} \quad (15)$$

where $\alpha \in \mathbb{R}^{2N \times d}$, and $\alpha_{\text{norm}}$ means L1 norm of $\alpha$. We use cosine similarity as default similarity. Because it has negative values, we adjust the interval to $[0, 1]$.

Therefore, we have a new attention-weighted contrastive loss to replace $L_{\text{cont}}$:

$$L_{W-\text{cont}} = -\log \frac{\exp(\omega_{i,j} \cdot \text{sim}(x_i, x_j)/\tau)}{\sum_{k=1}^{2N} 1_{\{k\}} \exp(\omega_{i,k} \cdot \text{sim}(x_i, x_k)/\tau)} \quad (16)$$

where $\omega_{i,j}$ indicates the element in the $i$-th row and $j$-th column of the attention matrix $W$, $1_{\{k\}}$ indicates that it is 1 when $k \neq i$, otherwise it is 0.

At this point, the total loss function can be defined as:

$$L = L_1 + \lambda_{\text{contrast}} L_{W-\text{cont}} \quad (17)$$

where $\lambda_{\text{contrast}}$ is also a hyper-parameter.

4. Experiments

We have implemented the DCCD framework based on the SNUNet [3] baseline and experiment it on two cross-domain change detection task between LEVIR-CD [21] dataset and WHU Building [22] dataset. While both datasets aim to facilitate change detection, they exhibit significant differences in their respective external environments. In this section, first, we introduce the datasets and experimental setup, including the implementation details. Subsequently, we introduce the evaluation metrics used in experiment. After that, the comparison between the performance of our framework and original method and the analysis of them is shown. Finally, we conduct an ablation experiment to demonstrate the effectiveness of the modules.
4.1. Datasets and Experimental Setup

Two change detection datasets is used in the experiments, which are LEVIR-CD [21] and WHU Building [22]. LEVIR-CD is a large scale remote sensing building change detection dataset, which includes 637 very high resolution (0.5 meter every pixel) google earth image patches pair. LEVIR-CD contains various buildings such as villas, high-rise apartments, small garages and large warehouses bitemporal image pair from 20 different regions in several cities in Texas, United States. The size of these image is 1024 × 1024. To accommodate the input size of model, 445 pairs of original picture are cropped to 256 × 256. As a result, 7120 patches is used in this experiment. WHU Building dataset contains original aerial data with 0.075m spatial resolution and 450km² covering in Christchurch, New Zealand and satellite imagery with about approximately 1m spatial resolution of buildings come from all over the world. Similarity, we also crop the size of image patches to 256 × 256 and consciously select image patches near the changing areas, as the proportion of changing areas in the WHU building dataset is relatively small. Finally, we use 7433 image patches in the experiment.

For the experiment setup, the server equipped with NVIDIA GeForce RTX 3090 (24G) GPU are used for deploying DCCD framework. Parameter initialization followed SNUNet original setup, while the dictionary and sparse representation matrix are initialized as 0 and 1 matrix respectively. Other framework parameters are initialized using the Kaiming initialization method [23]. For the data augmentation of original image, this work randomly combines cropping, image rotation, image scaling, image inversion and image translation as data augment. After that, the image becomes the input of network. We utilized the Adam optimizer with a learning rate set to 5e-4.

4.2. Evaluation Metrics

We use three evaluation metrics to investigate the performance of method: Precision (Pre), Recall (Rec) and F1 score. They can defined as:

\[
Pre = \frac{TP}{TP + FP} \quad (18)
\]

\[
Rec = \frac{TP}{TP + FN} \quad (19)
\]

\[
F1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \quad (20)
\]

where TP, FP and FN represent true positive, false positive and false negative, respectively.

4.3. Experiment Result and Analysis

The result of LEVIR-CD→WHU Building experiment is shown in Table[1], that is to say, we use the LEVIR-CD as train dataset and then test on WHU Building dataset with no label. “SNUNet-Source Only” means the original algorithm. “DCCD-SNUNet-ndcl” means the SNUNet combined with Student-Teacher network without the Dictionary Attention-Weighted contrastive learning module. “DCCD-SNUNet-ndl” indicates the DCCD method without dictionary learning attention module. And the “DCCD-SNUNet” represents the final version of DCCD framework combined with SNUNet.

The best performance has been bolded. It is obvious that as the module increases, the F1 score is also gradually increasing. Because F1 score combines the first two metrics, we believe
Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection

Table 1: Experiment Result in LEVIR-CD→WHU Building Cross-domain Task

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pre(%)</th>
<th>Rec(%)</th>
<th>F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNUNet-Source Only</td>
<td>40.48</td>
<td>76.79</td>
<td>53.01</td>
</tr>
<tr>
<td>DCCD-SNUNet-ndc1</td>
<td>54.80</td>
<td>66.61</td>
<td>60.13</td>
</tr>
<tr>
<td>DCCD-SNUNet-ndl</td>
<td>66.80</td>
<td>62.71</td>
<td>64.69</td>
</tr>
<tr>
<td>DCCD-SNUNet</td>
<td>69.47</td>
<td>61.11</td>
<td>65.02</td>
</tr>
</tbody>
</table>

that the F1 score is more representative of overall performance from all metrics. The F1 score of our method exceeds the original SNUNet nearly 12 percentage points, which demonstrates the method of reducing domain differences in this work is effective and our work can improve the performance of change detection method in cross-domain task. It also can be seen in Table 1 that the precision of DCCD is 28.99% higher than original SNUNet with Student-Teacher network increasing by 14.32%, contrastive learning module increasing by 12% and dictionary learning attention module increasing by 14.32%. For the recall, original SNUNet has the best performance, while DCCD-SNUNet decreased by 15.68%.

In addition, we also conduct some fine-tuning experiment to simply determine some hyper-parameters. For the temperature coefficient $\tau$, the result is shown in Table 2. It can be seen that temperature coefficient can not be too small, so we set it to 1 as default in the implementation.

Table 2: The performance of DCCD-SNUNet with different Temperature Coefficient

<table>
<thead>
<tr>
<th>Temperature Coefficient</th>
<th>Pre(%)</th>
<th>Rec(%)</th>
<th>F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>45.73</td>
<td>81.50</td>
<td>58.58</td>
</tr>
<tr>
<td>1</td>
<td>69.47</td>
<td>61.11</td>
<td>65.02</td>
</tr>
</tbody>
</table>

Besides the temperature coefficient, batch size is also an important hyper-parameters should be considered in experiment. We set different batch size to investigate the suitable batch size. Result of different batch size methods is shown in Table 3. The best performance is also bolded. It can be seen that 9 batch size has the best F1 performance than other. Although it has higher recall when batch size is 6, its precision is too low. As a result, we select that there are 9 image pairs in one batch as default. In the real implementation, the input contains 9 source domain image pairs, template image pairs and 9 target image pairs. However, due to the use of three GPUs in training, only one-third of the input images were used to calculate the loss function in reality. Thus, the $N$ in Dictionary Attention-Weighted module equals to one-third of 9, namely 3, as default. It is noted that if the number of GPU is changed, the actual batch size may be also changed.

Table 3: The performance of DCCD-SNUNet with different Batch Size

<table>
<thead>
<tr>
<th>Batch Size</th>
<th>Pre(%)</th>
<th>Rec(%)</th>
<th>F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>46.80</td>
<td>80.83</td>
<td>59.28</td>
</tr>
<tr>
<td>9</td>
<td>69.47</td>
<td>61.11</td>
<td>65.02</td>
</tr>
<tr>
<td>15</td>
<td>40.92</td>
<td>67.85</td>
<td>51.05</td>
</tr>
</tbody>
</table>
Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection

We also compare the time cost and parameter count of the proposed method, which is shown in Table 4. This result indicates that the parameter count of DCCD is slightly more than twice that of the original SNUNet, while the time consumption is nearly two-thirds of it. The parameter count is easier to understand because the average teacher framework uses two networks with the same structure. The shorter time consumption of DCCD is mainly due to the fact that we found DCCD typically converges within 50 epochs, which is fewer than the epochs required by SNUNet, making DCCD faster overall. If roughly estimated by comparing the training time per epoch, DCCD is slower than SNUNet. This is its limitation, as adding modules to the baseline increases the training time compared to the baseline itself.

Table 4: The Comparison of Parameter Count and Time Cost

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter count (KB)</th>
<th>Time cost (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNUNet-Source Only</td>
<td>47136</td>
<td>12.55</td>
</tr>
<tr>
<td>DCCD-SNUNet</td>
<td>105861</td>
<td>9.32</td>
</tr>
</tbody>
</table>

The visual comparisons of DCCD and its baseline on LEVIR-CD to WHU-Building dataset is shown in Fig.2. It is evident that SNUNet trained on LEVIR-CD tends to determine many background regions as change areas in prediction task on WHU-Building, whereas DCCD is less aggressive in this aspect. This is because DCCD enhances its generalization ability during the integration of both domains.

Figure 2: Visual Comparisons of DCCD and its baseline method on LEVIR-CD to WHU-Building task.
5. Conclusion and Future Work

In this paper, we propose a framework designed for unsupervised cross-domain remote sensing image change detection. Due to external factors such as season variations, sensor differences, illumination changes, and cloud occlusions, significant domain differences typically exist among various datasets. This discrepancy often results in considerable performance degradation when a model trained on one dataset is applied to another. Another challenge is the high cost associated with manual annotation, making it difficult to obtain a large number of labeled remote sensing images. These are two challenges we are concerned about in this study.

The proposed framework can be integrated with other models to improve the performance of it. The main ideas of this study include reducing the domain difference between source domain and target domains, and utilizing the Student-Teacher network to generate pseudo-label to address the issue of the study on dataset without labels. There are two phases in DCCD, the first is mixed domain construction phase and the second is alignment phase. In the experiments, the proposed method acquires better performance than state-of-the-art model, which demonstrate the effectiveness of it.

For DCCD, there is still a gap in performance compared to supervised algorithms. In the future work, a more effective mixing method need to be explored. And considering the dictionary attention-weighted contrastive module in this work is relatively simple, more complex attention mechanisms can be added to it in the future.

Author Contributions

Wei, T.: Methodology, Software, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization.

Liu, Y.: Formal analysis, Supervision.

Zhao, R.: Data Curation, Project administration, Funding acquisition.

Acknowledgements

The authors would like to thank the anonymous reviewers for their useful comments and suggestions which helped to improve the quality of this paper. This work was supported in part by National Natural Science Foundation of China under Grants 62106081 and 42301376, and in part by Project 2042024fg0013 supported by “the Fundamental Research Funds for the Central Universities”.

References

Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection


Author Biography

Tian Wei is a graduate student at Wuhan University, majoring in computer vision and domain adaptation. At present, his main research interests are source-free domain adaptation, image augmentation and contrastive learning.

Email: waitwei@whu.edu.cn; ORCID: https://orcid.org/0009-0002-6962-9408

Youfa Liu is an Associate Researcher in Wuhan University, Wuhan, China. He has published some papers in top journals, such as IJCV and IEEE TIP. He serves as a reviewer for top conferences and journals, such as AAAI, IEEE TIP, IEEE GRSL, Neural Networks, Pattern Recognition, Neurocomputing and Neural Processing Letters. His research interests include machine learning and computer vision.

Email: liuyfa1991@whu.edu.cn; ORCID: https://orcid.org/0000-0002-3540-5775

Rui Zhao received the B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2012, and the Ph. D. degree in photogrammetry and remote sensing...
Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection

from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2017. He is currently a Lecturer in the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include remote sensing, image processing, artificial intelligence, pattern recognition, deep learning and machine learning.

Email: zhaorui@nbu.edu.cn; ORCID: https://orcid.org/0000-0002-9577-7714