DCENet: Diff-Feature Contrast Enhancement Network for Semi-Supervised Hyperspectral Change Detection

Fulin Luo, *Senior Member, IEEE*, Tianyuan Zhou, Jiamin Liu, Tan Guo, *Member, IEEE*, Xiuwen Gong, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— Multitemporal hyperspectral images (HSIs) have wide applications in change detection (CD) of different land covers for their rich spectral features and image details. Traditional supervised learning-based HSI CD algorithms often rely on a substantial number of labeled samples. However, it requires a significant cost in sample annotation. In this article, we propose a diff-feature contrast enhancement network (DCENet) for semi-supervised HSI CD, which leverages a limited number of labeled samples to guide the training process and a large number of unlabeled samples to improve the confidence of CD. To achieve this, a differential fusion attention (DFA) subnetwork is constructed to extract temporal features from the initial input HSI patches. The dual-branch Siamese enhancement module (SEM) is utilized to enhance the generalization of differential features in the feature maps. Herein, multiscale Kullback–Leibler (KL) divergence and feature-enhanced probabilistic contrast loss are designed to constrain the SEM. The proposed method excels at detecting subtle changes in bitemporal HSIs simultaneously improving the generalization performance of networks. The visual and quantitative experimental results on four HSI datasets show that the proposed DCENet outperforms the compared state-of-the-art methods for HSI CD. Codes are available at https://github.com/Zhoutya/ChangeDetection-DCENet.**

***Index Terms*— Change detection (CD), contrastive learning, hyperspectral image, multiscale feature, Siamese network.**

I. Introduction

CHANGE detection (CD) based on remote sensing data is a crucial technique for detecting changes on the Earth's surface. It has a wide range of applications in urban planning, environmental monitoring, agriculture investigation, disaster assessment, and map revision [1]. CD aims to identify land-cover changes via two different periods of remote sensing images from the same area. Multitemporal hyperspectral remote sensing images (HSIs) are widely used for CD with continuous and detailed spectral features in a large electromagnetic wave range [2], [3].

Based on multitemporal HSIs, many classic CD methods have been developed by using the spectral information to construct a certain algebraic operation [4], such as image difference, image ratio, image regression, absolute distance, and change vector analysis (CVA) [5]. Other transformation-based methods project HSIs into a low-dimensional feature space that reveals the changed properties, which mainly includes principal component analysis (PCA) [6], independent component analysis (ICA) [7], and multivariate CD (MAD) [8]. These CD methods are often based on the spectral differences between different temporal HSIs, which cannot fully exploit the inherent characteristics of complex HSIs. With the advent of the convolutional neural network (CNN), more and more studies have embraced deep learning for the stronger adaptive feature extraction ability [9], [10]. For example, Ou et al. [11] constructed a self-supervised comparative pretraining model and trained the downstream CD network using high-confidence pseudo-labeled samples. Hu et al. [12] constructed a contrast learning network named HyperNet by constraining the consistency pairs of invariant pixels between the bitemporal images to learn invariant features. They belong to the unsupervised methods, which do not require any labeled data and focus on extracting intrinsic features from the data. Due to the lack of supervisory information, the accuracy of unsupervised methods is often limited, thereby typically falling short of the accuracy of supervised methods.

To utilize the supervisory information, supervised methods are developed with large amounts of labeled data for model training. Lin et al. [13] proposed a bilinear CNN (BCNN), establishing the relationship between bitemporal feature maps through the combined bilinear features. Mou et al. [14] and Chen et al. [15] proposed networks named ReCNN and SiamCRNN, using recurrent neural network (RNN) and long short-term memory (LSTM) [16] to capture the spatiotemporal relationship of bitemporal HSIs, respectively. Qu et al. [17] proposed a multilevel encoder–decoder attention network (MLEDAN), which introduces multiscale connection and

Manuscript received 2 November 2023; revised 11 January 2024; accepted 21 February 2024. Date of publication 7 March 2024; date of current version 21 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62071340, Grant 62371076, and Grant 62201109; in part by the Natural Science Foundation of Chongqing under Grant CSTB2022NSCQ-MSX0452; and in part by the Fundamental Research Funds for the Central Universities under Grant 2023CDJXY-039. *(Corresponding author: Jiamin Liu.)*

Fulin Luo and Tianyuan Zhou are with the College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: luoflyn@163.com; zhoutianyuan1016@163.com).

Jiamin Liu is with the Key Laboratory of Optoelectronic Technology and Systems, Education Ministry of China, Chongqing University, Chongqing 400044, China (e-mail: liujm@cqu.edu.cn).

Tan Guo is with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: guot@cqupt.edu.cn).

Xiuwen Gong is with the Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: xiuwen.gong@sydney.edu.au).

Xinbo Gao is with Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaoxb@cqupt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3374600

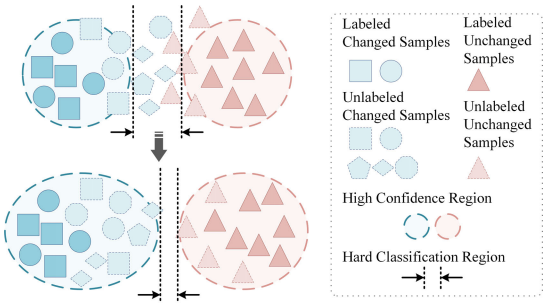


Fig. 1. Example to explain the improvement of confidence region with unlabeled samples.

attention mechanism to extract more effective spatial–spectral features. Luo et al. [18] proposed a multiscale diff-changed feature fusion network (MSDFFN), which focuses on multiscale extraction and fusion of changing features. However, these deep learning-based CD methods heavily rely on the accurate supervision with sufficient labeled samples for training, and their performance tends to decline as the size of training samples decreasing. In real, the acquisition of labeled samples is a time-consuming and labor-intensive process. Therefore, reducing the labeled samples to construct models has become a key issue in HSI CD.

In recent years, several studies have developed with limited labeled samples [19]. Qu et al. [20] proposed a dual-branch difference amplification graph convolutional network (D2AGCN), employing a graph CNN to process superpixels with a few training samples. Song et al. [21] proposed a cross-temporal symmetric attention network (CSA-Net), which extracts and integrates the joint spatial–spectral–temporal features by combining self-attention and CNN. Hu et al. [22] proposed an efficient multitemporal self-attention network (EMS-Net), and a supervised contrast loss was designed based on limited training samples to enhance the tightness of invariant detection. Hu et al. [23] constructed a global time-space interaction self-attention network named GlobalMind, which captures global information between samples in the entire graph by the multilevel multihead interaction attention. Dong et al. [24] constructed a graph transformer to represent the relationships between graph nodes. Wang et al. [25] proposed a joint spectral, spatial, and temporal transformer named SSTFormer to extract spectral sequence information and spatial texture information. With limited labeling information, these methods train model with very small samples. There are a lot of unlabeled samples in HSIs. To make full use of unlabeled samples, semi-supervised learning (SSL) has achieved significant success [26].

SSL utilizes a limited number of labeled samples and a large number of unlabeled samples to train model [27]. As shown in Fig. 1, the blue and pink symbols represent changed and unchanged samples, respectively, whereas light-colored symbols indicate unlabeled samples. In the field of CD, the changed samples generally tend to change as different land-cover categories, which results in the complex and diverse changed features compared with the unchanged features. The shaded regions in blue and pink depict the model's high-confidence regions to the changed and unchanged samples, and samples within these regions yield predictions with high confidence. When training involves only a small number of labeled samples, due to the limitation of the sample types, it is difficult for the model to establish accurate classification boundaries. With the introduction of unlabeled samples, the model's confidence region expands, reducing the hard classification regions. Noteworthy, unlabeled samples may play a significant role when the labeled samples are sparse and insufficient to cover all patterns in data.

Numerous studies have explored the integration of SSL in CD. Gong et al. [28] introduced a spectral and spatial attention network (S2AN), where a semi-supervised strategy was proposed by combining supervised and unsupervised methods to augment labeled training data. Zhao et al. [29] proposed a simplified 3-D convolution-based self-encoder (S3DCAECD) to extract spatial–spectral features using unsupervised learning, after which a classifier is trained with limited labeled samples to fine-tune the model. Liu et al. [30] proposed a multilayer cascade screening strategy (MCS4CD), which utilizes neighborhood spatial information and active learning to select highly reliable unlabeled samples for augmenting the training set. Chen et al. [31] adopt a margin sampling query function to progressively select high-confidence samples from the test set as additional training data.

The aforementioned studies aim to overcome the challenges posed by limited labeled data and expand the generalization ability of model. However, supervised learning-based methods are often limited by the size of labeled samples. When the labeled sample size is small, complex supervised models are often difficult to train and prone to overfitting. On the other hand, some semi-supervised methods often rely on adding traditional preprocessing methods to generate pseudo-labels of unlabeled samples, which does not take full advantage of the data properties of unlabeled data itself. Some other semi-supervised methods need to train a pretrained model with unlabeled samples and fine-tune it with a small number of labeled samples. Obviously, these methods involve two stages, which are inefficient and unable to achieve end-to-end learning.

Motivated by these observations and insights, we propose a diff-feature contrast enhancement network (DCENet) for semi-supervised HSI CD. As shown in Fig. 2, it can learn representative and general change features from bitemporary HSIs. DCENet is composed of a differential fusion attention (DFA) subnetwork and a Siamese enhancement module (SEM). The DFA subnetwork with a multilevel res-attention (MRA) block is designed to extract differential features from HSI patches. The SEM is proposed to enhance the differences and invariant features in the differential feature maps. In addition, we introduce a combination of multiscale Kullback–Leibler (KL) divergences and feature-enhanced probabilistic contrast loss. This combination loss enforces the consistency between the Siamese branches, thus mining the intrinsic information of the unlabeled samples and reinforcing the models' ability to extract robust differential information. The main contributions of this article include the following.

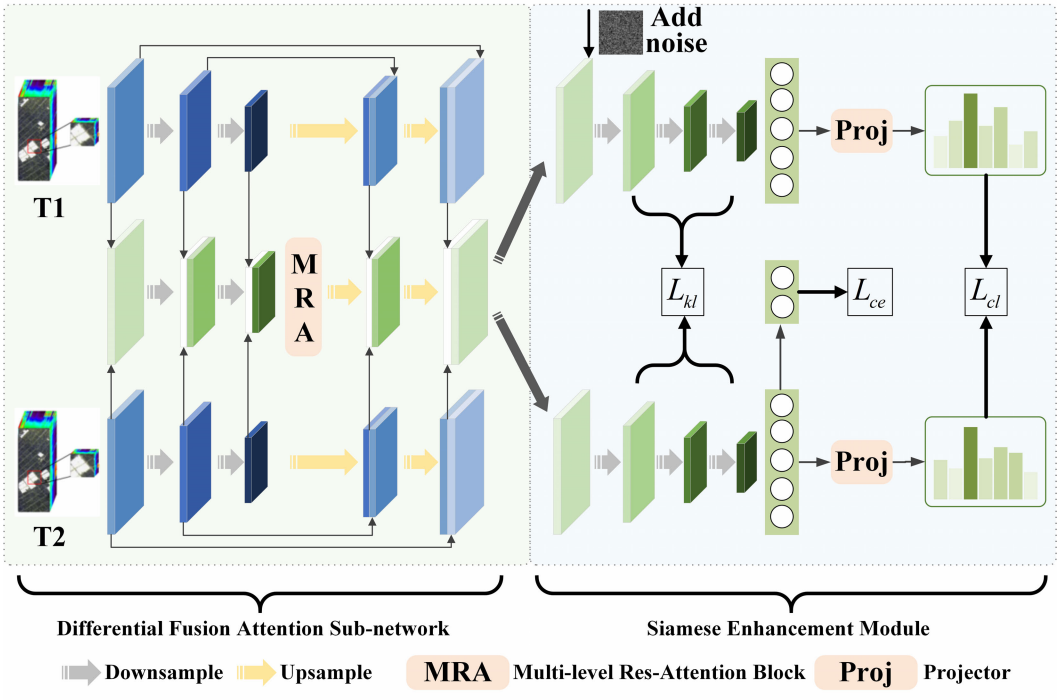


Fig. 2. Overview of the proposed DCENet for HSI CD.

1) We design a DFA subnetwork to extract the features from multitemporal HSIs, where MRA is developed to aggregate attention and enhance the representation of change information.
2) The proposed multiscale KL divergence constraints the consistency of multiscale feature maps between the noise-added and noise-free Siamese branches, facilitating the model to learn more generalized features.
3) The feature-enhanced probabilistic contrast loss is proposed to mine the change semantic consistency of features. This loss function employs probabilities in place of features in the conventional contrast loss, thereby enhancing the compactness of each class.
4) Based on the above three thoughts, we propose an end-to-end DCENet. DCENet uses limited labeled samples and a large number of unlabeled samples to train the model efficiently and does not need additional predetection and pretraining. The experimental results on four HSI datasets show its significant performance.

The rest of this article is organized as follows. Section II introduces the details of the proposed DCENet. Section III presents the experiments. In the end, Section IV draws some conclusions of this article and suggestions for future work.

II. Proposed Method

In this section, we introduce the proposed DCENet for the semi-supervised HSI CD task in Fig. 2, which is composed of the DFA and the SEM. The bitemporal HSI patches are passed through the DFA subnetwork to get the differential feature map. A multihead attention fusion module is designed in the DFA subnetwork to enhance the local variance information. Then, diff-changed features go through the SEM, and the consistency of the Siamese branches is constrained by a multiscale KL divergences constraint and a contrast constraint to obtain a model with better generalization performance. In the following, we will explain each module of the network in detail.

A. DFA Subnetwork

1) Architecture: An encoder–decoder network [32] is widely used in CD tasks due to its excellent feature extraction ability. In the DFA subnetwork, we construct three codec branches: two branches for bitemporal HSI patches and a branch for the diff-feature maps. As shown in Fig. 2, the bitemporal HSIs are fed into the upper and lower temporal codec branches to obtain rich multiscale features. Subsequently, the diff-feature maps are obtained by subtracting feature maps of the same scale from the bitemporal feature maps. These multiscale differential features are then integrated into the intermediate differential feature codec branch to assist network in learning.

The two temporal branches include the encoding path, decoding path, and two skip connection operations. The encoder extracts features with a series of downsampling based on convolution operations, and the decoder recovers the resolution by continuously upsampling based on deconvolution operations. Given an initial feature map $F \in \mathbb{R}^{C\times H\times W}$ as the input patch, the encoder stage of the two temporal branches can be summarized as follows:

$$E_{T_t}^{i} = \mathrm{ReLU}\left(f_{\mathrm{Conv}}{}^{3\times 3}\left(E_{T_t}^{i-1}\right)\right), \quad i = 1, 2, 3 \tag{1}$$

where $t = 1, 2$ represents two different temporal, $E_{T_t}^{i}$ represents the feature map generated by the ith convolutional layer for the input patch, and $f_{\mathrm{Conv}}{}^{3\times 3}$ represents a convolution

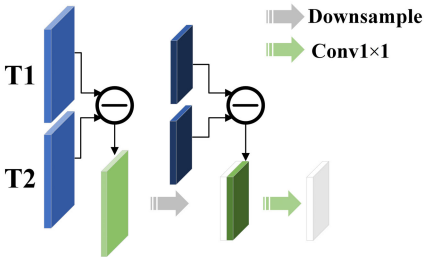


Fig. 3. Details of differential operation in DFA subnetwork.

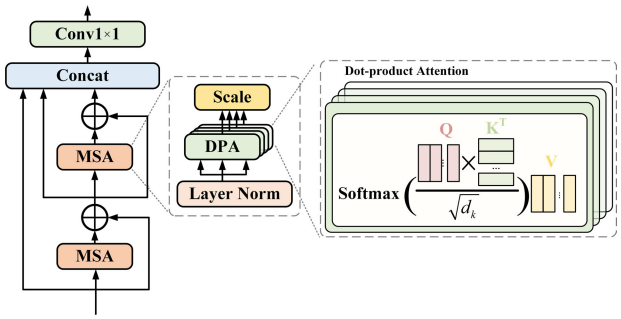


Fig. 4. Structure of the proposed MRA block.

operation with a kernel size of 3 × 3. All convolution layers are followed by a batch normalization layer and a rectified linear unit (ReLU) layer. ReLU(·) denotes the ReLU activation function.

In the encoding path, the previous feature map is sampled by a convolution operation without padding and then passing through the ReLU activation function and outputting to the next layer. With successive downsampling operations, the model can learn more abstract features from HSI patches step by step.

The decoder stage of the temporal branches can also be summarized as follows:

$$D_{T_t}^i = \begin{cases} f_{\text{Conv}}{}^{1\times1}\left[\text{ReLU}\left(f_{\text{DConv}}{}^{3\times3}\left(E_{T_t}^{i+2}\right)\right); E_{T_t}^{i+1}\right], & i=1 \\ f_{\text{Conv}}{}^{1\times1}\left[\text{ReLU}\left(f_{\text{DConv}}{}^{3\times3}\left(D_{T_t}^{i-1}\right)\right); E_{T_t}^{i-1}\right], & i=2 \end{cases} \tag{2}$$

where $D_{T_t}^i$ represents the output feature map after the ith deconvolution operation, $f_{\text{DConv}}{}^{3\times3}$ represents a deconvolution operation with a kernel size of 3 × 3, and $f_{\text{Conv}}{}^{1\times1}$ represents a convolution operation with a kernel size of 1 × 1. [;] represents the stacking operation along the channel dimension, which achieves the skip connections of feature maps. By incorporating skip connections between multilayer downsampling and upsampling paths, we can mitigate the loss of fine-grained details. Noteworthy, the bitemporal encoding and decoding branches share the network weights [33].

The detailed operation between the bitemporal branches and the differential branch is shown in Fig. 3. Initially, the feature maps of the bitemporal network are subjected to subtraction at corresponding scales to yield the differential feature maps, which can be expressed as follows:

$$\begin{aligned} E_d^i &= E_{T_1}^i - E_{T_2}^i, \quad i=1,2,3 \\ D_d^i &= D_{T_1}^i - D_{T_2}^i, \quad i=1,2 \end{aligned} \tag{3}$$

where E_d^i and D_d^i represent the differential feature maps.

Subsequently, the initial feature map obtained from the first subtraction operation serves as the start of the differential branch. The differential branch includes the encoding path, the decoding path, and the MRA block. The coding path can be summarized as follows:

$$E_D^i = \begin{cases} E_d^i, i=1 \\ f_{\text{Conv}}{}^{1\times1}\left[\text{ReLU}\left(f_{\text{Conv}}{}^{3\times3}\left(E_D^{i-1}\right)\right); E_d^{i-1}\right], & i=2,3 \end{cases} \tag{4}$$

where E_D^i represents the feature map generated by the ith convolutional layer of the input patches. The differential maps obtained from the bitemporal branches are fused by stacking along the channel. The decoder stage can also be summarized as follows:

$$D_D^i = \begin{cases} f_{\text{Conv}}{}^{1\times1}\left[\text{ReLU}\left(f_{\text{DConv}}{}^{3\times3}\left(\text{MRA}\left(E_D^{i+2}\right)\right)\right); D_d^i\right], & \\ \qquad\qquad\qquad\qquad\qquad\qquad i=1 & \\ f_{\text{Conv}}{}^{1\times1}\left[\text{ReLU}\left(f_{\text{DConv}}{}^{3\times3}\left(D_D^{i-1}\right)\right); D_d^i\right], \quad i=2 & \end{cases} \tag{5}$$

where D_D^i represents the output feature map after the ith deconvolution operation and MRA(·) denotes the proposed MRA block.

2) MRA Block: To focus on key information, attention mechanism is proposed in CNN and is widely used in the field of deep learning [34]. The idea of attention mechanism is to learn a set of weighting coefficients by the network autonomously and dynamically emphasize the regions of interest with the weights [35]. In this DFA subnetwork, we propose an MRA block at the bottleneck to adaptively filter out redundant features, focusing on the change information in the feature map, as shown in Fig. 4.

The core module is scaled dot-product attention, which adaptively learns long-range relationships between individual elements within the feature maps. Meanwhile, the multihead enables the output of attention layer to contain information about coded representations across different subspaces and enhances the expressive power of model. The weight vector is subsequently computed using Q, K, and V as follows:

$$\text{DPA}(Q,K,V) = \prod_h \left[\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V\right]_h \tag{6}$$

where Q, K, and V are obtained from the input by layer normalization and three different linear layers. h is the number of multihead, and the value is set to be 4 in the MRA block. $\sqrt{d}$ is a scaling factor that prevents the value from being too large or too small. softmax(·) denotes the softmax function, which can be described as follows:

$$\text{softmax}(x) = \frac{\exp(x_i)}{\sum_{j=1}^k \exp(x_j)}. \tag{7}$$

Inspired by the feature pyramid architecture [36], we design an MRA block based on the MSA layer. The input feature map $x \in \mathbb{R}^{C\times H\times W}$ goes through two layers of MSA to obtain two levels of attention maps. These two multilevel attention maps

are concatenated together with the original inputs x and fused by 1×1 convolution to produce the output. We introduce the residual connection to the MSA layer for preserving the original features [37]. The specific process can be summarized as follows:

$$\mathrm{MRA}(x) = f_{\mathrm{Conv}}{}^{1\times1}[x; \mathrm{MSA}(x); \mathrm{MSA}(\mathrm{MSA}(x))] \quad (8)$$

where $\mathrm{MRA}(x) \in \mathbb{R}^{C\times H\times W}$ is the final output. The multilevel attention integrates varying levels of attention information, enabling the network to learn more characteristics associated with change. When dealing with shallow features containing complex fine-grained information, the attention may bring in a lot of redundant information. For the deep semantic information in the decoding stage, the use of multihead attention will introduce more computational complexity, which will actually affect the decoding of features. Therefore, we use the MRA block in the bottleneck of the differential encoder–decoder branch to best integrate the semantic information, which optimally integrates the advanced semantic information while mitigating the computational burden.

B. Siamese Enhancement Module

The current applications of unlabeled samples in SSL remote sensing CD are mainly of the following two types. One type is to generate pseudo-labels for unlabeled samples using traditional unsupervised algorithms. These high-confidence pseudo-labels are used as supervised information for training. The other type is to use unlabeled samples to construct an unsupervised pretrained model while fine-tuning it with supervised information in downstream tasks. To simplify the preprocessing and postprocessing process and achieve end-to-end CD model training, we design an SEM to perform unsupervised learning on unlabeled samples to extract their intrinsic features for training and enhance the generalization performance of the model to unknown samples.

1) Architecture: Contrastive learning is widely used in unsupervised and self-supervised models [38]. The idea of contrastive is that, for a robust model, the output should be approximated even if the input is perturbed [39]. To enhance the robustness of the differential information in the feature maps and to achieve unsupervised learning on unlabeled samples, we construct a Siamese structure based on a noise-added branch and an original branch for contrastive learning.

After the input bitemporal HSI patches pass through the DFA subnetwork, we can obtain feature maps D_D^2. We consider the original feature maps as $F_O{}^0$. Define a noise that obeys the Gaussian distribution of mean μ and variance σ, denoted as $x \sim N(\mu, \sigma^2)$. The noise addition strategy can improve the generalization ability of the model to noise. We add noise pixel-by-pixel and get a feature map as $F_N{}^0$. Next, for the obtained noise-added feature map $F_N{}^0$ and the original feature map $F_O{}^0$, we perform multiple downsampling to extract multiscale features, respectively. The operation can be summarized as follows:

$$\begin{aligned} F_O^i &= \mathrm{ReLU}\left(f_{\mathrm{Conv}}{}^{3\times3}\left(F^{i-1}\right)\right), \quad i = 1, 2, 3 \\ F_N^i &= \mathrm{ReLU}\left(f_{\mathrm{Conv}}{}^{3\times3}\left(F_N^{i-1}\right)\right), \quad i = 1, 2, 3 \end{aligned} \quad (9)$$

where F_O^i and F_N^i represent the feature map from the ith convolutional layer in the noise-added branch and the original branch, respectively. Next, we design two loss functions to constrain the consistency of the two branches, so that the network can learn more robust change features.

2) Multiscale KL Divergence: KL divergence is a similarity metric of two probability distributions, which is widely used as a classical loss function in machine learning tasks [40]. KL divergence measures how much information is lost when approximating one distribution to another distribution. Suppose that for a random variable ξ, there exist two probability distributions P and Q, defining the KL divergence from P to Q as follows:

$$\mathrm{KL}(P, Q) = \sum_{\xi=1}^{n} P(\xi) \ln\left(\frac{P(\xi)}{Q(\xi)}\right) \quad (10)$$

where $\mathrm{KL}(\cdot, \cdot)$ denotes the calculation of KL divergence.

In SEM, after three downsamplings, we obtain feature maps of different scales. The scales of ground object features are diverse in HSIs. Therefore, feature maps of different scales may focus on ground objects with various sizes. To efficiently strengthen the correlation between the Siamese branches, achieving multidimensional constraints across various scales, we design a multiscale KL divergence loss. Thus, we can obtain a KL divergence value from each layer of the SEM and sum all the KL divergence values to combine the patch–patch correlations of the multilayer features into a more informative metric as follows:

$$L_{\mathrm{KL}} = \sum_{k=1}^{K} \left(\mathrm{KL}\left(F_O{}^k, F_N{}^k\right)\right) \quad (11)$$

where $K = 3$ denotes the number of downsampling operations and $\mathrm{KL}(\cdot, \cdot) \in \mathbb{R}^N$ means the KL divergence between discrete distributions of the multiscale feature maps from the two branches. Compared with the single-scale KL scatter, the designed multiscale KL divergence simultaneously constrains the consistency of the feature maps at different scales.

3) Feature-Enhanced Probabilistic Contrast Loss: The standard contrastive learning acts on the extracted features with $L2$ normalization [41]. For his CD, we find that contrastive learning with the standard paradigm does not perform well. In CD, unchanging trends are consistent, while changing trends are of various types. Applying consistency constraints directly to the bitemporal HSIs promotes the learning of invariant features [42]. However, for change samples, the spectral representations at corresponding positions in bitemporal HSIs are different, and blindly aligning samples at these positions is not reasonable. Therefore, we apply noise-added consistency contrastive learning to the differential feature maps, aiming to simultaneously enhance both the change features and the invariant features. Based on this, we design a probabilistic contrastive loss function to constrain the Siamese branches'

outputs based on cosine similarity, enhancing the features' robustness against noise.

First, a projector is designed to facilitate the spatial and spectral features to obtain the embedding vectors. The projector consists of two dense 1×1 convolutional layers with a BN layer and a ReLU activation function

$$z = \mathrm{MLP}(F^3) = \mathrm{BN}({f_{\mathrm{Conv}}}^{1\times1}(\mathrm{ReLU}({f_{\mathrm{Conv}}}^{1\times1}(F^3)))) \quad (12)$$

where BN denotes the batch normalization operation.

Given $z_O \in \mathbb{R}^C$ from the original path and $z_N \in \mathbb{R}^C$ from the noise-added path, the proposed loss function is defined as follows:

$$L_{\mathrm{CL}} = -(2 - \cos\_) \otimes \cos\_ \quad (13)$$

$$\cos\_ = \cos(P(z_O), \quad \mathrm{StopGrad}(P(z_N))) \quad (14)$$

where $P(\cdot)$ denotes the probability. StopGrad refers to halting the gradient backpropagation of the current branch, only receiving gradient information from another branch, playing a key role in the success of contrastive learning in Siamese networks. $\cos(\cdot, \cdot)$ represents the calculation of cosine similarity and can be defined as follows:

$$\cos(z_O, z_N) = \frac{z_O}{\|z_O\|_2} \cdot \frac{z_N}{\|z_N\|_2}. \quad (15)$$

In (13), $(2 - \cos\_)$ acts as an adjustment factor in the cosine loss. For difficult-to-distinguish samples, the more similar the two samples are, the closer to 1 the cosine similarity is. This adjustment factor enhances the weights of hard samples, compelling the model to pay more attention to distinguishing these challenging samples.

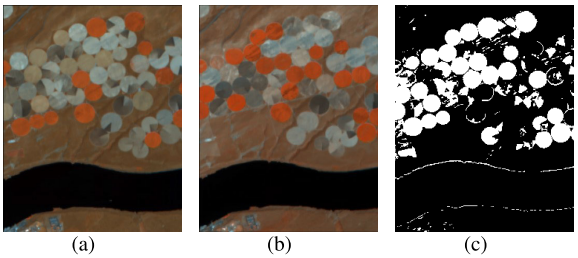
Another crucial design is probability contrastive. When using features to calculate contrastive loss, similar sample pairs that are brought closer might be quite scattered in the feature space. Utilizing probabilities for the calculation can promote a more consistent and compact semantic representation within each respective class. Specifically, using the softmax to transform the features into a probability distribution, and transform the representation from a feature level to a semantic level. Compared with sharpening operations, employing softmax is considerably more gentle and nuanced. This method promotes more adaptive and nuanced alignment of features, allowing for more accurate representation and distinction of intrinsic characteristics within each class.

C. Loss Function

1) CE Loss: After the DFA subnetwork and SEM, we obtain a feature map with rich change details. We choose the feature map from the third downsampling of the original noise-free branch for the final detection. The probability estimate obtained from the fully connected layers can be used to predict the final labels of the input patches. The final CD results can be described as follows:

$$y_p = \mathrm{softmax}\left(fc\left(F_O^3\right)\right) \quad (16)$$

where y_p is the predicted probability result and F_O^3 is the output after the third downsampling in noise-free branches. The fc denotes the fully connection layers to extract the features and reduce dimension.



(a) (b) (c)

Fig. 5. Hermiston dataset. (a) Image acquired in 2013. (b) Image acquired in 2014. (c) Ground truth.

CD task can be considered as a binary classification task, and each pixel of bitemporary HSIs is divided into two categories, i.e., change and unchange. Therefore, we supervise the limited labeled samples using the cross-entropy loss function, and the loss function is calculated as follows:

$$L_{\mathrm{ce}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log y_p + (1 - y_i)\log(1 - y_p)) \quad (17)$$

where n denotes the number of samples and y_i is the ground-truth label of the given sample.

2) Overall Loss: In the training strategy, labeled samples and unlabeled samples are sequentially passed through the network and backpropagated. Thus, we compute the loss function separately and add them together as the overall loss function, and the formulas are shown below

$$L_{\mathrm{cl}} = L_{\mathrm{CL}}^{\mathrm{label}} + L_{\mathrm{CL}}^{\mathrm{unlabel}} \quad (18)$$

$$L_{\mathrm{kl}} = L_{\mathrm{KL}}^{\mathrm{label}} + L_{\mathrm{KL}}^{\mathrm{unlabel}}. \quad (19)$$

The total loss function consists of the above three loss functions as follows:

$$L_{\mathrm{total}} = L_{\mathrm{ce}} + L_{\mathrm{cl}} + L_{\mathrm{kl}}. \quad (20)$$

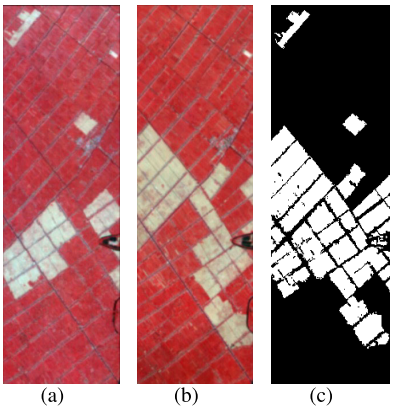
In the training stage, the model is optimized by minimizing the loss function using gradient back propagation. In the testing stage, the output of the noise-free branch is used for testing.

III. Experimental Results and Analysis

In this section, we first describe the HSI-CD datasets and the evaluation metrics to validate the effectiveness of the model. Then, we briefly describe the corresponding comparison algorithms and specific experimental details. Then, a series of comparison experiments are provided to validate the model effects, as well as ablation experiments to verify the effectiveness of each module.

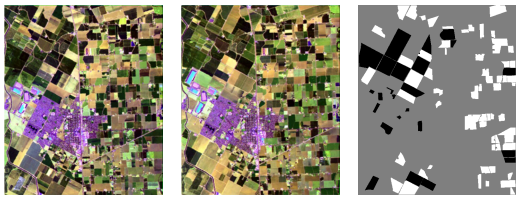
A. Datasets and Evaluation Measures

1) Datasets: The first dataset, named "Hermiston," as shown in Fig. 5, belongs to an irrigated farmland from the Hermiston City, USA, which was acquired in 2013 and 2014. This dataset was obtained by the Hyperion sensor mounted on the EO-1 satellite. The spatial size of each image is 307×241 pixels including 154 spectral bands after eliminating noise. The main change is farmland cover.



(a) (b) (c)

Fig. 6. Farmland dataset. (a) Image acquired on May 3, 2006. (b) Image acquired on April 23, 2007. (c) Ground truth.



(a) (b) (c)

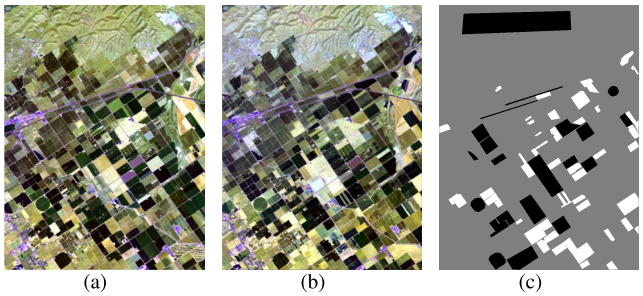
Fig. 7. Bay dataset. (a) Image acquired in 2013. (b) Image acquired in 2015. (c) Ground truth.

The second dataset, named "Farmland," belongs to a farmland near the city of Yancheng, Jiangsu, China, which was acquired by Earth Observing-1 (EO-1) Hyperion sensor on May 3, 2006, and April 23, 2007, respectively. The dataset has 242 bands in the range of 0.4–2.5 m with a spatial resolution of 30 m, as shown in Fig. 6. After removing noise and water absorption bands, it contains 155 spectral bands for experiments, and the spatial size of each image is 450 × 140 pixels. The main change areas are farmland.

The third dataset, named "Bay," as shown in Fig. 7, was taken in 2013 and 2015 with the AVIRIS sensor surrounding the city of Patterson (California). Bay dataset has a large spatial size as 600 × 500 pixels and 224 spectral bands. The main change areas are covered by farmlands and buildings. Note that there are a large number of unknown regions, and only the labeled changed and unchanged areas are adopted for training and assessment.

The fourth dataset, named "Barbara," exhibited as Fig. 8, was shot in the years 2013 and 2014 with the AVIRIS sensor in the Santa Barbara region. The spatial dimensions are 984 × 740 pixels and both have 224 spectral bands. The two HSIs have recorded the urban evolution and dynamic changes of farmland.

2) Evaluation Measures: To better quantify the performance of the proposed method, we mainly used the overall accuracy (OA) and Kappa coefficient (KC) as metrics; precision (Pr), recall (Re), and $F1$ score ($F1$) were introduced as an auxiliary evaluation. When the proportion of samples from different categories is very unbalanced, the categories with large proportion often have a large influence on OA to evaluate the effectiveness of model. So, we introduced Pr, Re, and their harmonic mean $F1$ as synthetic evaluation.



(a) (b) (c)

Fig. 8. Barbara dataset. (a) Image acquired on 2013. (b) Image acquired on 2014. (c) Ground truth.

The metrics are defined as follows:

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \tag{21}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{22}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{23}$$

$$F1 = \frac{2\mathrm{PR}}{\mathrm{P}+R} \tag{24}$$

$$\mathrm{Kappa} = \frac{\mathrm{OA}-p_e}{1-p_e} \tag{25}$$

$$p_e = \frac{(\mathrm{TP}\times\mathrm{FN})+(\mathrm{TP}\times\mathrm{FP})+(\mathrm{TN}\times\mathrm{FN})+(\mathrm{TN}\times\mathrm{FP})}{N^2} \tag{26}$$

where true positive (TP) indicates the number of pixels correctly classified as changed region, true negative (TN) denotes the number of pixels correctly classified as unchanged regions, false positive (FP) represents the number of pixels misclassified as changed regions, and false negative (FN) is the number of pixels misclassified as unchanged regions. The larger value of these evaluation metrics indicates better detection performance. To show the experimental results more prominently, TP, TN, FP, and FN are shown in white, red, green, and black in the visualization maps, respectively.

B. Compared Methods and Experimental Details

To evaluate the performance of the proposed architecture, we further compared our method with some other CD methods. We classify the comparison models into three categories.

1) Unsupervised Models: The first category is based on unsupervised models without labeled training samples, including CVA [5], PCAKM [6], and HyperNet [12]. CVA is the most commonly used traditional method, which can detect changes by change intensity and change direction. PCAKM uses the PCA method to project the original data into a new lower dimensional feature space, and the CD is achieved by k-means clustering. HyperNet constructed a contrast learning without negative sample pairs to learn invariant features from bitemporary images.

2) Large-Sample-Based Supervised Models: The second category is supervised models originally based on larger amounts of labeled training samples (such as 20% of the training samples), which we refer to as large-sample supervised models. Representative algorithms include BCNNs [13],

MLEDAN [17], and MSDFFN [18]. BCNNs [13] find the relationship between bitemporal feature maps by combining bilinear features. MLEDAN [17] learns the discriminating features by introducing multiscale features, and LSTM was introduced to construct the correlation between the bitemporal phases. MSDFFN [18] focuses on differential feature learning and designs a multiscale differential feature fusion module.

3) Small-Sample-Based Models: The third category is supervised models designed for a small number of training samples, including S2AN [28], MCS4CD [30], D2AGCN [20], GlobalMind [23], CSA-Net [21], EMS-Net [22], and DLIEG [24]. S2AN [28] introduces an S2AN, wherein they combine real sparse label mapping to augment the labeled samples. MCS4CD [30] proposes a strategy for selecting reliable unlabeled samples using neighborhood spatial information and active learning. S2AN and MCS4CD are implemented based on SSL and can perform better on a small-sample size. D2AGCN [20] constructs a graphical CNN for the full map to mine long-range connections between pixels. GlobalMind [23] uses a global multilevel multihead interaction attention to capture global information between samples in the entire graph. CSA-Net [21] extracts and integrates the joint spatial–spectral–temporal features by combining self-attention and CNN. EMS-Net [22] designs a supervised contrast loss to enhance the tightness of invariant detection. The last approach, DLIEG [24], constructs a graph transformer to model the relationships between sequences of graph nodes. Most of these methods use a full graph input to allow information from all samples to participate in network learning, reducing the reliance on labeled samples.

4) Experimental Details: In our network, comprehensively considering the complexity of calculation and spatial–spectral information, we chose the input patch size as 7×7. Our network was trained and tested on an NVIDIA GeForce 3090 GPU with 24G memory using the PyTorch [43] framework. As for large-sample-based supervised models, we reproduced the codes of these comparison algorithms based on the original paper.

In experiments, we selected 0.2% labeled samples from the datasets and 10%, 10%, 5%, and 10% unlabeled samples for Farmland, Hermiston, Bay, and Barbara datasets as the training samples, the rest as the testing samples, where the amount of unlabeled samples is based on experimentally determined. We selected the training samples by using 4, 3, 1, and 3 as random seeds for the Farmland, Hermiston, Bay, and Barbara datasets, respectively. In the stage of training, we used the SGD optimizer [44] with a weight decay of $5e-3$. The initial learning rate was designed to be $1e-1$ and decayed by a factor of 0.1 at every 30 epoch. The number of total epochs was 100, the batch size of labeled samples was set to 64, and the batch size of unlabeled samples was also set to 64.

As for small-sample-based models, due to the difficulty of reproduction, and considering the differences in the experimental settings, we chose our training samples to the same level and directly compare them with the experimental results from the original paper. However, due to different datasets for each paper, it is difficult for us to find all consistent comparison algorithms. So, we selected the corresponding comparison algorithms for each dataset within the available options. A batch size of 64 is still used when the labeled sample size is less than 1%, and a batch size of 128 is used when the labeled sample size is higher than 1%.

C. Experimental Results

1) Comparison With Unsupervised and Large-Sample Supervised Models: Table I shows the results of each model on the four datasets. In comparison with traditional CVA and PCAKM methods, the supervised learning methods yield better precision and have a great improvement in terms of KCs, which indicates that there is a good consistency between the predicted change of the model and the actual change results. It also indicates that the CVA and PCAKM algorithms misjudge a large number of invariant regions into changing regions, thus having a high Re but a low KC. Deep learning-based methods, including BCNNs, MLEDAN, and MSDFFN, generally outperform unsupervised methods due to their ability to learn more intricate features through convolutional layers. MLEDAN and MSDFFN have a better performance compared with BCNNS. On the Farmland dataset, the simple CVA algorithm can also obtain an accuracy of 95.25%, indicating the ease of partitioning of the dataset itself. On this dataset, MLEDAN obtains a better performance than MSDFFN, while on the other datasets, MLEDAN performs poorly compared with MSDFFN, especially on the Bay dataset, where the MLEDAN algorithm suffers from severe overfitting with a labeled sample size of 0.2%. Compared with all the methods, the proposed DCENet model has the best performance in OA and KC, respectively. The designed semi-supervised strategy can effectively use the information from the unlabeled samples to help the network learn better features.

Figs. 9 and 10 show the visualization results of the experiments on the Farmland and Hermiston datasets. From the visual observations, compared with the other methods, our proposed DCENet presents the fewest FP pixels, thus achieving the best visual performance. For the Farmland dataset, the unsupervised CVA and PCAKM methods exhibit more misclassified pixels, with significant "salt and pepper" noise in the unchanged areas (black regions), and a large number of misclassification pixels around the edges of the changed areas and small targets, because they are unsupervised algorithms and lack enough learning of spatial–spectral information. The deep learning algorithms, such as BCNNs, MLEDAN, and MSDFFN, have better performance in distinguishing unchanged pixels; however, the areas between the rice fields, as shown in the edges of the block areas of the image, still appear some pixels of FN. The DCENet has fewer misclassification points and better CD details than MLEDAN and MSDFFN, mainly shown in the middle-right small-target areas of the image.

For the Hermiston dataset, the change region is presented as a circular region with partially connection edges, the edges of the circular region can be easily misclassified, and some of the independently scattered regions can be easily ignored. Compared with the other methods, our proposed DCENet preserves as many target change regions as possible and reduces the loss of independent change regions. For the

TABLE I
COMPARISONS OF DCENET WITH THE UNSUPERVISED AND LARGE-SAMPLE SUPERVISED METHODS ON THE FOUR DATASETS

Dataset	Methods	OA%	KC	F1%	Pr%	Re%
Farmland	CVA[5]	95.25	0.8860	91.97	90.33	93.66
	PCAKM[6]	95.14	0.8837	91.82	89.78	93.96
	HyperNet[12]	/	/	/	/	/
	BCNNs[13]	94.30	0.8632	90.36	88.70	92.09
	MLEDAN[17]	96.12	0.9060	93.53	92.79	93.91
	MSDFFN[18]	95.57	0.8949	92.65	89.33	**96.24**
	ours	**96.56**	**0.9156**	**93.97**	**95.72**	92.27
Hermiston	CVA[5]	92.02	0.7416	78.85	97.90	66.01
	PCAKM[6]	92.01	0.7413	78.83	**97.90**	65.98
	HyperNet[12]	92.06	0.7613	81.12	87.40	75.69
	BCNNs[13]	89.85	0.6896	75.26	83.53	68.49
	MLEDAN[17]	90.35	0.6962	75.47	88.29	65.91
	MSDFFN[18]	92.86	0.7834	82.80	90.59	76.25
	ours	**94.62**	**0.8389**	**87.29**	93.40	**81.93**
Bay	CVA[5]	82.55	0.6558	83.00	94.16	74.20
	PCAKM[6]	82.81	0.6606	83.32	94.05	74.79
	HyperNet[12]	90.79	0.8152	91.29	92.24	90.37
	BCNNs[13]	92.74	0.8519	93.64	94.24	93.04
	MLEDAN[17]	86.96	0.7347	88.49	89.70	87.32
	MSDFFN[18]	95.01	0.8985	95.58	97.14	94.08
	ours	**95.58**	**0.9099**	**96.10**	**97.30**	**94.93**
Barbara	CVA[5]	83.24	0.6515	79.15	77.47	80.90
	PCAKM[6]	83.21	0.6517	79.24	77.11	81.50
	HyperNet[12]	91.14	0.8148	88.80	88.36	89.25
	BCNNs[13]	96.05	0.9169	94.92	96.17	93.70
	MLEDAN[17]	95.86	0.9129	94.68	95.73	93.65
	MSDFFN[18]	96.36	0.9232	95.27	**97.54**	93.10
	ours	**97.85**	**0.9549**	**97.26**	97.51	**97.00**

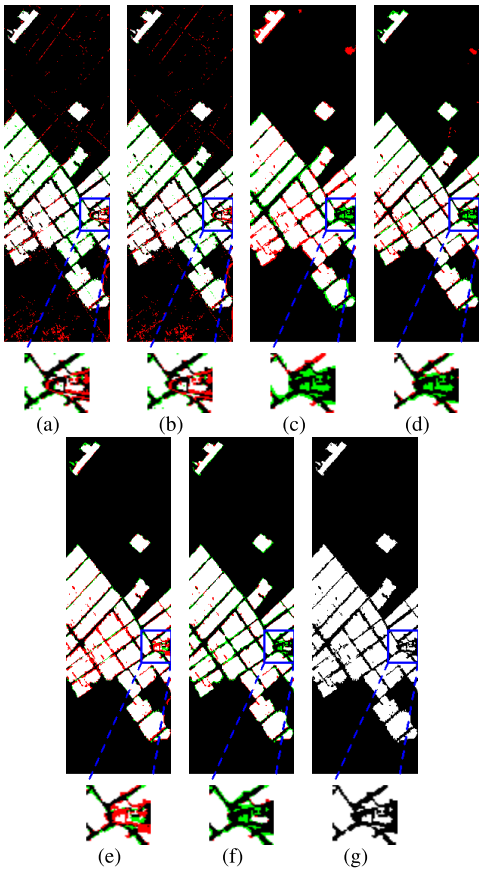


Fig. 9. Visualized results of different methods on the Farmland dataset. (a) CVA, (b) PCAKM, (c) BCNNs, (d) MLEDAN, (e) MSDFFN, (f) DCENet, and (g) ground truth.

detection results of CVA and PCAKM, a lap of unchanged pixels around the circular change areas are misclassified into changes. The supervised methods, such as BCNNs, MLEDAN, and MSDFFN, show a lot of FN pixels, and some large circular regions are lost in their CD results. Compared with MSDFFN, DCENet has fewer pixels of FN and FP and shows better CD details.

Figs. 11 and 12 show the visualization results of the experiments on the Bay and Barbara datasets. These two datasets contain a large number of unlabeled samples, shown in gray in the figure. Compared with the other methods, CVA and PCAKM have more misclassified pixel points, such as the misclassified red point in the black block, which means the misclassification of unchanged pixels into changed pixels. Another phenomenon is the green point in the white block which means the misclassification of changed pixels into unchanged pixels. The deep learning-based methods have fewer misclassified pixel points, but some large regions are still lost and misclassified on the region edges. Compared with the other algorithms, our model can detect all large regions and has better visualization performance.

2) Comparison With Small-Sample-Based Models: The proposed DCENet is based on a small number of labeled samples of 0.2%, so a comparison with some other small-sample models is necessary. Due to the difficulty of reproducing the original algorithm and the objective factors of the experimental environment, we chose to the same sample size for comparison. Four comparison algorithms were selected for each

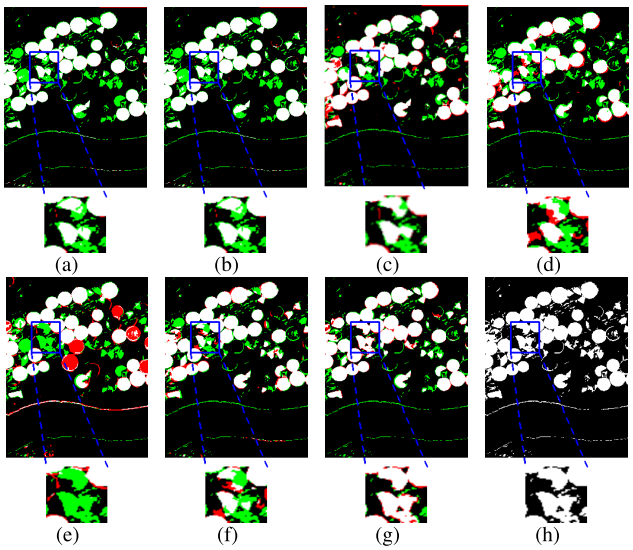


Fig. 10. Visualized results of different methods on the Hermiston dataset. (a) CVA, (b) PCAKM, (c) HyperNet, (d) BCNNs, (e) MLEDAN, (f) MSDFFN, (g) DCENet, and (h) ground truth.

dataset except the Hermiston dataset. The reason is that, so far, we have only found one model named GlobalMind, which has utilized the same Hermiston dataset as ours.

The experimental results are shown in Table II. For the Farmland dataset, the proposed DCENet outperforms all the

TABLE II
COMPARISON DCENET WITH THE SMALL-SAMPLE-BASED MODELS ON THE FOUR DATASETS

Dataset	Methods	samples(%)	samples(n)	OA%	KC	F1%	Pr%	Re%
Farmland	D2AGCN[20]	1.59	1000	93.74	0.8568	/	/	/
	MCS4CD[30]	0.32	200	91.59	0.8013	/	/	/
	CSA-Net[22]	0.32	200	94.22	0.8641	90.56	86.09	95.52
	EMS-Net[22]	0.32	200	95.60	0.8961	92.76	97.28	88.65
	DLIEG[24]	0.5	315	92.70	0.8316	88.59	/	/
	ours	**0.2**	**126**	**96.56**	**0.92**	**93.97**	**95.72**	**92.27**
Hermiston	GlobalMind[23]	1.35	998	95.56	0.8781	90.71	96.20	85.82
	ours			**95.72**	**0.8728**	**89.99**	**95.11**	**85.39**
	MCS4CD[30]	0.27	200	94.46	0.8458	/	/	/
	ours	0.2	147	**94.62**	**0.8389**	**87.29**	**93.40**	**81.93**
Bay	D2AGCN[20]	1.52	1000	96.90	0.9376	/	/	/
	GlobalMind[23]	1.36	893	98.15	0.9629	98.26	97.58	98.94
	ours			**98.49**	**0.9692**	**98.68**	**99.17**	**98.30**
	DLIEG[24]	0.5	328	96.05	0.9087	93.76	/	/
	ours			**97.17**	**0.9421**	**97.54**	**97.44**	**97.63**
	CSA-Net[22]	0.3	200	95.68	0.9135	95.88	93.95	97.89
	ours			**96.33**	**0.9252**	**96.78**	**97.42**	**96.15**
Barbara	D2AGCN[20]	0.754	1000	98.03	0.9588	/	/	/
	GlobalMind[23]	0.75	994	98.65	0.9717	98.28	98.06	98.51
	ours			**98.86**	**0.9762**	**98.56**	**98.12**	**99.01**
	DLIEG[24]	0.5	663	96.54	0.9286	97.06	/	/
	ours			**98.53**	**0.9693**	**98.14**	**97.92**	**98.37**
	S2AN[28]	0.07	93	93.46	0.8621	91.52	/	/
	ours			**94.21**	**0.8781**	**92.54**	**93.86**	**91.26**

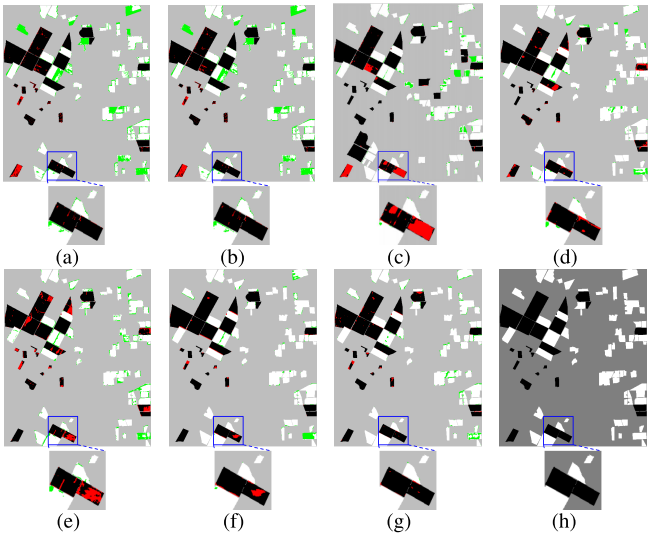


Fig. 11. Visualized results of different methods on the Bay dataset. (a) CVA, (b) PCAKM, (c) HyperNet, (d) BCNNs, (e) MLEDAN, (f) MSDFFN, (g) DCENet, and (h) ground truth.

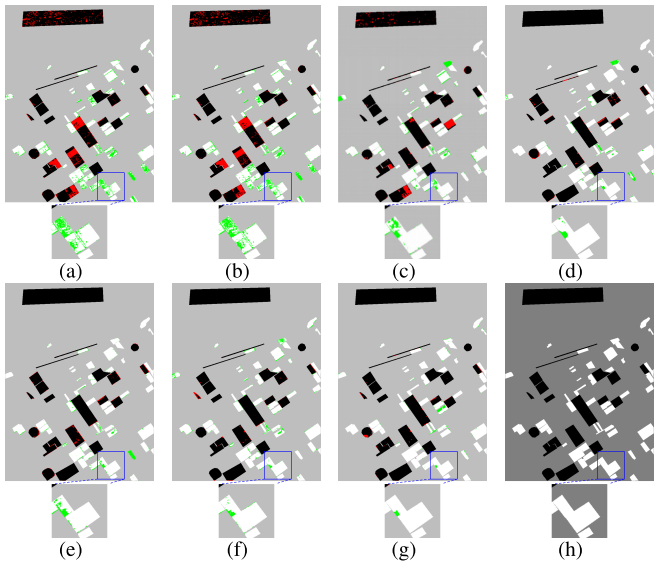


Fig. 12. Visualized results of different methods on the Barbara dataset. (a) CVA, (b) PCAKM, (c) HyperNet, (d) BCNNs, (e) MLEDAN, (f) MSDFFN, (g) DCENet, and (h) ground truth.

other models only on 0.2% of labeled training samples. Due to the easy differentiation of the Farmland dataset itself, the efficient feature extraction ability and the introduction of unlabeled samples of our model are key to achieving this accuracy. For the Hermiston dataset, we used the sample size of 1.35% to compare with GlobalMind, and our model has a 0.1% improvement in OA accuracy compared with GlobalMind. As for the Bay dataset and Barbara dataset, when increasing the labeled sample size to the same as the comparison algorithm, our DCENet is higher than all the other models. When decreasing the labeled sample size to 0.7% for comparison with S2AN, our method is still higher, indicating that the proposed DCENet still has an advantage on small-sample sizes.

D. Ablation Study

1) Ablation of Each Module: In this article, we proposed some modules to improve the performance of change feature learning. To more clearly show the effectiveness of each module, we conducted the ablation experiments for each module, including MRA block, multiscale KL divergence,

TABLE III
COMPARISONS ABLATION OF EACH MODULE ON FOUR DATASETS

Dataset	Model					
	unlabelled	×	✓	✓	✓	✓
	MRA	×	×	✓	✓	✓
	CLloss	×	×	×	✓	✓
	KLloss	×	×	×	×	✓
Hermiston	OA%	92.94	93.64	94.16	94.51	**94.62**
	KC	0.7796	0.8124	0.8281	0.8356	**0.8389**
	F1%	82.28	85.28	86.53	87.03	**87.29**
	Pr%	**94.65**	89.22	90.15	93.12	93.40
	Re%	72.77	81.67	**83.19**	81.68	81.93
Farmland	OA%	95.66	96.06	96.37	**96.60**	96.56
	KC	0.8958	0.9064	0.9124	**0.9175**	0.9156
	F1%	92.66	93.45	93.80	**94.15**	93.97
	Pr%	90.86	90.21	92.91	93.80	**95.72**
	Re%	94.54	**96.94**	94.72	94.51	92.27
Bay	OA%	93.74	94.52	94.56	95.25	**95.58**
	KC	0.8736	0.8890	0.8896	0.9033	**0.9099**
	F1%	94.35	95.09	95.15	95.80	**96.10**
	Pr%	**97.97**	97.83	97.46	97.20	97.30
	Re%	90.99	92.50	92.94	94.44	**94.93**
Barbara	OA%	97.18	97.27	97.44	97.51	**97.85**
	KC	0.9407	0.9424	0.9463	0.9478	**0.9549**
	F1%	96.36	96.47	96.73	96.83	**97.26**
	Pr%	97.99	**98.25**	97.41	97.16	97.51
	Re%	94.79	94.75	96.06	96.49	**97.00**

TABLE IV
ABLATION OF THE L_{cl} ON THE HERMISTON AND BAY DATASETS

Dataset	Model	OA%	KC	F1%	Pr%	Re%
Hermiston	only w/ labeled	92.99	0.7857	82.93	91.89	75.56
	only w/ unlabeled	91.51	0.7260	77.61	**95.59**	65.33
	w/o softmax	94.17	0.8293	86.65	89.58	**83.91**
	w/o proj	94.21	0.8234	85.95	94.89	78.54
	ours	**94.62**	**0.8389**	**87.29**	93.40	81.93
Bay	only w/ labeled	95.46	0.9077	95.98	97.72	94.29
	only w/ unlabeled	95.16	0.9017	95.71	97.41	94.08
	w/o softmax	95.08	0.8996	95.70	96.04	**95.37**
	w/o proj	95.29	0.9043	95.81	**97.91**	93.79
	ours	**95.58**	**0.9099**	**96.10**	97.30	94.93

TABLE V
ABLATION OF THE L_{kl} ON THE HERMISTON AND BAY DATASETS

Dataset	Model	OA%	KC	F1%	Pr%	Re%
Hermiston	only w/ labeled	93.75	0.8117	85.09	92.09	79.08
	only w/ unlabeled	93.48	0.8035	84.44	91.38	78.48
	F^3	94.49	0.8331	86.75	**94.78**	79.98
	F^2+F^3	94.39	0.8327	86.81	92.40	81.87
	F^1+F^2+F^3	**94.62**	**0.8389**	**87.29**	93.40	**81.93**
Bay	only w/ labeled	94.74	0.8925	95.40	95.69	95.12
	only w/ unlabeled	94.73	0.8929	95.33	96.98	93.75
	F^3	95.17	0.9016	95.76	96.65	94.89
	F^2+F^3	94.57	0.8892	95.23	95.97	94.51
	F^1+F^2+F^3	**95.58**	**0.9099**	**96.10**	**97.30**	**94.93**

and feature-enhanced probabilistic contrast loss on the four datasets. Specifically, we added modules step by step and designed five experiments. The first experiment is the baseline, which does not add any unlabeled sample. Table III shows the experimental results.

By analyzing the experimental results on the four datasets, with the gradual addition of the proposed modules, the accuracies are improved compared with the previous model in general. For the Farmland dataset, the addition of multiscale KL divergence shows a very slight decrease in model accuracy, but on the other datasets, the addition of the multiscale KL divergence loss results in an improvement of 0.1%–0.3%. Therefore, we also retain the module as our point of innovation. The reason for this on the Farmland dataset is that the dataset itself is easy to divide, and the addition of the KL loss makes the network more complex and easier to overfit. For the other three datasets, the complete model presents optimal values in OA and Kappa. In general, the $F1$ score is more appropriate to evaluate the model for unbalanced classification problem, and the $F1$ score of our proposed DCENet is still optimal. In summary, although the proposed modules show side effects in a few cases for some datasets, they generally have the advantage to improve the performance of CD under most conditions for all the experimental datasets.

2) Ablation of L_{cl}: Next, we take the Hermiston and Bay datasets, for example, to discuss the details for specific modules. First, we compare the performance of the proposed method using different contrastive structures to determine the advantage of feature-enhanced probabilistic contrast loss. Experimental results are shown in Table IV. To ensure the fairness of the comparison, we used the same settings in all the experiments.

According to the results, both constructing contrast loss only for labeled samples and only for unlabeled samples have difficulty in achieving optimal accuracy. This suggests that introducing unlabeled samples is effective. Meanwhile, adding effective constraints on unlabeled samples to mine information and facilitate training is essential. From the experimental results on both datasets, the accuracy decreases when the softmax operation or projection is removed. The softmax operation transforms the feature consistency into the semantic consistency, achieving more accurate consistency constraints on different change types and allowing the network to better detect the changed pixels. The projection head maps the encoded representations into the potential space for the contrast loss, adaptively removing some of the redundant information and making the retained features more favorable for consistency constraints. In summary, the proposed feature-enhanced probabilistic contrast loss can bring the greatest enhancement to the network.

3) Ablation of L_{kl}: In the proposed DCENet framework, the feature scale is an inevitable parameter in our multiscale KL divergence. In this article, we used feature maps for all three scales of 1, 3, and 5 to calculate the overall L_{kl}. We tried to select different scales to verify the effectiveness of the multiscale strategy. The results are shown in Table V.

First, we can see that the results are consistent with the L_{cl}; when training only labeled samples or unlabeled samples, neither of them can achieve the highest accuracy. It proves the effectiveness of introducing unlabeled samples to participate in training. Second, for the KL loss, the experimental results show that the optimal results occur when the overall KL

TABLE VI
ABLATION OF THE MRA BLOCK ON THE HERMISTON AND BAY DATASETS

Dataset	Model	OA%	KC	F1%	Pr%	Re%
Hermiston	w/o attention	94.37	0.8347	87.06	90.39	**83.96**
	w/ ECA	94.26	0.8290	86.53	91.77	81.87
	w/ CBAM	94.43	0.8338	86.90	92.51	81.93
	w/ MSA	94.21	0.8257	86.21	93.10	80.27
	MRA block	**94.62**	**0.8389**	**87.29**	**93.40**	81.93
Bay	w/o attention	95.07	0.9000	95.60	**98.02**	93.29
	w/ ECA	95.47	0.9079	95.99	97.64	94.40
	w/ CBAM	95.44	0.9072	95.96	97.58	94.40
	w/ MSA	95.16	0.9016	95.71	97.35	94.13
	MRA block	**95.58**	**0.9099**	**96.10**	97.30	**94.93**

TABLE VII
DISCUSSION OF THE NOISE ADDITION STRATEGIES ON THE HERMISTON AND BAY DATASETS

Dataset	Noise types and scales	OA%	KC	F1%	Pr%	Re%
Hermiston	multiplicative(1)	94.25	0.8266	86.27	**93.42**	80.14
	additive(0.05)	93.74	0.8096	84.87	93.12	77.97
	additive(0.2)	94.23	0.8282	86.47	91.78	81.73
	additive(0.5)	93.89	0.8157	85.39	92.61	79.23
	ours(0.1)	**94.62**	**0.8389**	**87.29**	93.40	**81.93**
Bay	multiplicative(1)	94.64	0.8913	95.22	**97.55**	93.01
	additive(0.05)	94.82	0.8945	95.42	96.81	94.07
	additive(0.2)	94.64	0.8907	95.30	95.94	94.67
	additive(0.5)	94.77	0.8936	95.39	96.63	94.18
	ours(0.1)	**95.58**	**0.9099**	**96.10**	97.30	**94.93**

divergence is calculated using the multiscale feature maps. The multiscale feature maps are more helpful for the overall consistency of the Siamese network and facilitates model learning to obtain more robust features.

4) Ablation of the MRA Block: In recent years, a lot of attention mechanisms have been used in computer vision. ECA [45] utilizes the adaptive 1-D convolution to replace the full connection layer, which simplifies the calculation. CBAM [46], cascading channel attention and spatial attention, uses max-pooling and average-pooling operations to aggregate spatial and channel information. To validate the effectiveness of the attention enhancement, we compared the proposed MRA block with the other classic attention, i.e., ECA, CBAM, and a single MSA. The results are shown in Table VI.

From the results, we can see that the proposed MRA block can yield the best accuracies and is more suitable for the CD task. The attention mechanisms, such as ECA and CBAM, focus on the information of channels and spaces. The proposed MRA block uses the multihead attention mechanism to capture long-range features that can better fuse global information. The MRA block learns deeper attentional information and gets more discriminating features by overlaying multiple levels of MSA compared with a single MSA.

E. Discussion

1) Discussion of the Noise Addition Strategies: In the proposed DCENet, we constructed the SEM to enhance the discriminative power of differential features. In the course of this research, the intensity of the noise added to the branches has a great impact on the final accuracy. Thus, we tried different classes and different intensities of noise to validate our model. The results are shown in Table VII.

We design a random Gaussian white noise with mean = 1 and variance = 0. We try to add the noise in the form of multiplication and addition, as well as multiplying the noise by different ratios or adding it. The value in parentheses indicates the magnitude of the added noise. Specifically, the original Gaussian noise is multiplied by this value before being added to the feature maps. From the experimental results, the network accuracy is optimal when the ratio of Gaussian white noise is 0.1. When the additive noise is too large, the difference between the two Siamese branches is too large, and the network becomes difficult to train. When the additive noise is too small, the difference between the two Siamese branches is too small, and then, the consistency constraint cannot urge the network to learn useful information.

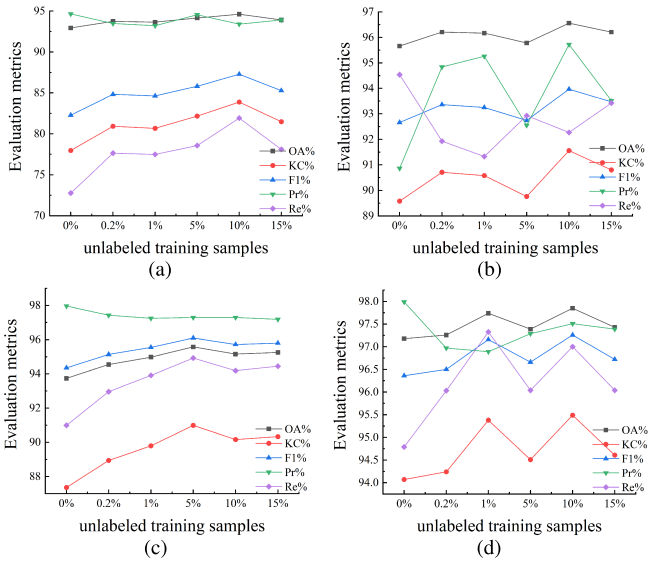


Fig. 13. CD results under different unlabeled sample sizes on four datasets. (a) Hermiston, (b) Farmland, (c) Bay, and (d) Barbara.

2) Discussion of the Unlabeled Training Sample Size: To comprehensively investigate the influence of the unlabeled training sample size on the detection results, we tested the detection results under different unlabeled training samples on the four datasets. We set the selected unlabeled training sample size as 0% (i.e., training with only labeled samples), 0.2%, 1%, 5%, 10%, and 15%. We presented the experimental results in a curve line chart. The OAs, KCs, and other evaluation indicators are shown in Fig. 13.

With the increasing of unlabeled sample size, the network accuracy fluctuates to various degrees. Based on the experimental results, we selected 10%, 10%, 5%, and 10% as the unlabeled samples from the Farmland, Hermiston, Bay, and Barbara datasets, respectively. A large amount of experience in supervised learning-based learning has shown that, in general, the more sample sizes for training are used, the better the network results will be. That is why various data augmentation strategies continue to emerge in the finite sample case. Current experiments show that adding unlabeled samples appropriately helps the model learn more generalized features,

improving its performance. However, increasing the number of unlabeled samples does not always improve accuracy. This occurs because increasing sample size has less impact on the network when most sample types are already covered by the training samples. At the same time, this indicates that SSL strategies are important, and how to balance the ratio between labeled and unlabeled samples is one of the issues that need further research.

IV. Conclusion

In this article, an end-to-end framework named DCENet, including DFA subnetwork and SEM, was proposed to detect the changed regions of bitemporal HSIs. The proposed DFA subnet, with a designed MRA block, can obtain a differential feature that contains a wealth of information from the input patch pairs. SEM uses multiscale KL divergence and feature-enhanced probabilistic contrast loss to constrain the two Siamese branches. Experimental results on four HSI datasets show that the proposed method can produce more accurate CD results than the other compared methods. There are also some limitations of our proposed method. Our method is a semi-supervised algorithm that needs labeled samples and necessitates a balance between labeled and unlabeled sample sizes. In future work, we plan to further develop unsupervised models to mine features from unlabeled samples.

References

[1] Z. Lv, M. Zhang, W. Sun, J. Atli Benediktsson, T. Lei, and N. Falco, "Spatial-contextual information utilization framework for land cover change detection with hyperspectral remote sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3336791.

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

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

[4] M. Hasanlou and S. T. Seydi, "Hyperspectral change detection: An experimental comparative study," *Int. J. Remote Sens.*, vol. 39, no. 20, pp. 7029–7083, Oct. 2018.

[5] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2007.

[6] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[7] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.

[8] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998.

[9] Z. Lv, P. Zhang, W. Sun, J. Atli Benediktsson, and T. Lei, "Novel land-cover classification approach with nonparametric sample augmentation for hyperspectral remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3309949.

[10] L. Hu, Q. Liu, J. Liu, and L. Xiao, "PRBCD-Net: Predict-refining-involved bidirectional contrastive difference network for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3314217.

[11] X. Ou, L. Liu, S. Tan, G. Zhang, W. Li, and B. Tu, "A hyperspectral image change detection framework with self-supervised contrastive learning pretrained model," *IEEE J. Sel. Topics Appl. Earth Obser. Remote Sens.*, vol. 15, pp. 7724–7740, Sep. 2022.

[12] M. Hu, C. Wu, and L. Zhang, "HyperNet: Self-supervised hyperspectral spatial–spectral feature understanding network for hyperspectral change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3218795.

[13] Y. Lin, S. Li, L. Fang, and P. Ghamisi, "Multispectral change detection with bilinear convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1757–1761, Oct. 2020.

[14] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral–spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[15] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[17] J. Qu, S. Hou, W. Dong, Y. Li, and W. Xie, "A multilevel encoder–decoder attention network for change detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3130122.

[18] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, "Multiscale diff-changed feature fusion network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3241097.

[19] Z. Lv, P. Zhang, W. Sun, T. Lei, J. Atli Benediktsson, and P. Li, "Sample iterative enhancement approach for improving classification performance of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[20] J. Qu, Y. Xu, W. Dong, Y. Li, and Q. Du, "Dual-branch difference amplification graph convolutional network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3135567.

[21] R. Song, W. Ni, W. Cheng, and X. Wang, "CSANet: Cross-temporal interaction symmetric attention network for hyperspectral image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[22] M. Hu, C. Wu, and B. Du, "EMS-Net: Efficient multi-temporal self-attention for hyperspectral change detection," 2023, *arXiv:2303.13753*.

[23] M. Hu, C. Wu, and L. Zhang, "GlobalMind: Global multi-head interactive self-attention network for hyperspectral change detection," 2023, *arXiv:2304.08687*.

[24] W. Dong, Y. Yang, J. Qu, S. Xiao, and Y. Li, "Local information-enhanced graph-transformer for hyperspectral image change detection with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023, Art. no. 5509814.

[25] Y. Wang et al., "Spectral–spatial–temporal transformers for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3203075.

[26] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2168–2187, Apr. 2022.

[27] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[28] M. Gong et al., "A spectral and spatial attention network for change detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3139077.

[29] C. Zhao, H. Cheng, and S. Feng, "A spectral–spatial change detection method based on simplified 3-D convolutional autoencoder for multitemporal hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[30] L. Liu, D. Hong, L. Ni, and L. Gao, "Multilayer cascade screening strategy for semi-supervised change detection in hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Obser. Remote Sens.*, vol. 15, pp. 1926–1940, Feb. 2022.

[31] Y. Chen, M. Zhu, C. Zhao, S. Feng, Y. Fan, and Y. Tang, "A hyperspectral change detection method based on active learning strategy," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 6700–6703.

[32] Z. Lv, H. Huang, W. Sun, T. Lei, J. Atli Benediktsson, and J. Li, "Novel enhanced UNet for change detection using multimodal remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[33] Z. Lv, J. Liu, W. Sun, T. Lei, J. Atli Benediktsson, and X. Jia, "Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3334521.

[34] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3279–3298, Apr. 2023.
[35] K. Jiang, J. Liu, W. Zhang, F. Liu, and L. Xiao, "MANet: An efficient multidimensional attention-aggregated network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3328334.
[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[38] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.
[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[40] T. van Erven and P. Harremos, "Rényi divergence and kullback–leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
[41] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.
[42] J. Li, Y. Zhang, Z. Wang, K. Tu, and S. Hou, "Probabilistic contrastive learning for domain adaptation," 2021, *arXiv:2111.06021*.
[43] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[44] M. Andrychowicz et al., "Learning to learn by gradient descent by gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
[45] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

Fulin Luo (Senior Member, IEEE) received the B.E. degree in mechanical engineering and automation from Southwest Petroleum University, Chengdu, China, in 2011, and the M.E. and Ph.D. degrees in instrument science and technology from Chongqing University, Chongqing, China, in 2013 and 2016, respectively.

He was an Associate Researcher and a Post-Doctoral Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, from 2017 to 2021. He was a Research Fellow with Nanyang Technological University, Singapore, from 2020 to 2021. He has been a Professor with the College of Computer Science, Chongqing University, since 2022. His research interests include remote sensing processing, computer vision, and biomedical analysis.

Tianyuan Zhou received the B.S. degree in electronic information engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2021. She is currently pursuing the Ph.D. degree with the College of Computer Science, Chongqing University, Chongqing, China.

Her research interests include hyperspectral image change detection, remote sensing image processing, and machine learning.

Jiamin Liu received the M.S. and Ph.D. degrees in instrument science and technology from Chongqing University, Chongqing, China, in 1998 and 2001, respectively.

He is currently an Associate Professor with Chongqing University. His research interests include biometrics, image processing, and pattern recognition in general.

Tan Guo (Member, IEEE) received the M.E. degree in signal and information processing and the Ph.D. degree in communication and information systems from Chongqing University, Chongqing, China, in 2014 and 2017, respectively.

He has been a Lecturer with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, since 2018. He was a Post-Doctoral Fellow with Macau University of Science and Technology, Macau, China, from 2020 to 2022. His research interests include computer vision, pattern recognition, and machine learning.

Xiuwen Gong received the B.E. and M.E. degrees from Anhui Normal University, Wuhu, China, in 2011 and 2014, respectively, and the Ph.D. degree in artificial intelligence from the Faculty of Engineering, The University of Sydney, Sydney, SA, Australia, in 2023.

She is a Research Fellow with The University of Sydney, since 2023. Her research interests include machine learning, deep learning, and artificial intelligence.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively.

From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University, Xi'an, China. He is currently a Professor of control science and engineering with Xidian University and the President of Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include computer vision, machine learning, pattern recognition, artificial intelligence, and smart city.